

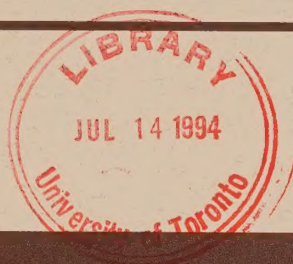
3 1761 10374382 9

12
-001



SURVEY METHODOLOGY

19



Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

June 1994

•

VOLUME 20

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1994 • VOLUME 20 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1994

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

June 1994

Price: Canada: \$45.00
United States: US\$50.00
Other Countries: US\$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members B.N. Chinnappa C. Patrick
G.J.C. Hole D. Roy
F. Mayda (Production Manager) M.P. Singh
R. Platek (Past Chairman)

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
M.J. Colledge, <i>Statistics Canada</i>	L.-P. Rivest, <i>Université Laval</i>
J.-C. Deville, <i>INSEE</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J.D. Drew, <i>Statistics Canada</i>	C.-E. Särndal, <i>Université de Montréal</i>
W.A. Fuller, <i>Iowa State University</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	J. Sedransk, <i>State University of New York</i>
D. Holt, <i>University of Southampton</i>	J. Waksberg, <i>Westat Inc.</i>
G. Kalton, <i>University of Michigan</i>	K.M. Wolter, <i>National Opinion Research Center</i>
A. Mason, <i>East-West Center</i>	A. Zaslavsky, <i>Harvard University</i>

Assistant Editors N. Laniel, M. Latouche, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue No. 12-001) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 20, Number 1, June 1994

CONTENTS

In This Issue	1
Small Area Estimation	
M.P. SINGH, J. GAMBINO and H.J. MANTEL	
Issues and Strategies for Small Area Data	3
Comment: W.A. FULLER	15
G. KALTON	18
Response from the Authors	21
D. HOLT and D.J. HOLMES	
Small Domain Estimation for Unequal Probability Survey Designs	23
A.C. SINGH, H.J. MANTEL and B.W. THOMAS	
Time Series EBLUPs for Small Areas Using Survey Data	33
<hr/>	
J.G. KOVAR and E.J. CHEN	
Jackknife Variance Estimation of Imputed Survey Data	45
D.S. TRACY and S.S. OSAHAN	
Estimation in Overlapping Clusters with Unknown Population Size	53
N.G.N. PRASAD and J.E. GRAHAM	
PPS Sampling over Two Occasions	59
R.R. SITTER and C.J. SKINNER	
Multi-way Stratification by Linear Programming	65
W.A. FULLER, M.M. LOUGHIN and H.D. BAKER	
Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey	75
E.A. STASNY, B.G. TOOMEY and R.J. FIRST	
Estimating the Rate of Rural Homelessness: A Study of Nonurban Ohio	87

Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743829>

In This Issue

This issue of *Survey Methodology* opens with a special section on **Small Area Estimation**. The three papers in this special section consider the problem of domain estimation from a variety of perspectives. I would like to give special thanks to Jon Rao for coordinating the editorial work for this special section. One or two other papers on this topic, which were not yet ready for publication, may also appear in a later issue.

The first paper in the special section, by Singh, Gambino and Mantel, considers the problem of small area statistics from the perspective of survey design. They discuss the role of sample design features such as stratification, clustering and sample allocation in the production of small area statistics for both planned and unplanned domains. A short overview of current approaches to small area estimation is also included. The paper is followed by insightful comments by Fuller and Kalton and a response from the authors.

The paper by Holt and Holmes presents a model based approach to small area estimation that does not “borrow strength” from other domains, and which may be used when auxiliary totals and means are not available. Estimates of model parameters are combined with design based estimates of means or totals of covariates. Using an example from market research it is shown that the method can lead to significant gains in efficiency of estimates for small domains.

The last paper in the special section, by Singh, Mantel and Thomas, presents an empirical comparison of several different small area estimators using simulated sampling from a population of farms. It is shown that, in the context of repeated surveys, estimators based on time series models can perform better, with respect to both bias and mean squared error, than those based on models for a single time point.

Kovar and Chen present results of a simulation study in which they investigated statistical properties of the jackknife approach to variance estimation of imputed data sets. Under this approach, the variance due to imputation is incorporated in the variance estimator. Real data sets, four different imputation methods, simple random sampling and a uniform nonresponse mechanism were used. Performance under a stratified multistage design and a non-uniform nonresponse mechanism was also studied.

Tracy and Osahan propose ratio estimators associated with two sampling strategies for estimation of a population mean in overlapping clusters with unknown population size. While much work by several researchers is available on non-overlapping clusters in the literature, there are many practical sampling situations where one gets overlapping clusters. The first sampling strategy is an equal probability with replacement sampling scheme while the second strategy is an unequal probability sampling scheme.

Prasad and Graham extend the “Random Group Method” for sampling with probability proportional to size (PPS) to sampling over two occasions. They use for this purpose the information on a study variate observed on the first occasion to select the matched portion of the sample on the second occasion.

Sitter and Skinner show how linear programming may be used to find an optimal sample design in the context of a multi-way stratification. Their approach is compared to existing methods both by illustrating the sampling schemes generated for specific examples and by evaluating mean squared errors. Variance estimation is also considered.

Fuller, Loughin and Baker consider regression weighting in the presence of non-response. They exhibit conditions under which the regression estimator remains consistent in the presence of non-response, and discuss implications for the choice of regressor variables. The ideas are illustrated by application to the 1987-88 Nationwide Food Consumption Survey conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture.

The paper by Stasny, Toomey and First gives a description of a survey conducted in 1990 to estimate the rate of rural homelessness in Ohio. The possible magnitude of the bias of the estimator is investigated by simulating sampling from a variety of synthetic populations. It is found that the bias is likely to be small compared to the standard deviation.

Issues and Strategies for Small Area Data

M.P. SINGH, J. GAMBINO and H.J. MANTEL¹

ABSTRACT

This paper identifies some technical issues in the provision of small area data derived from censuses, administrative records and surveys. Although the issues are of a general nature, they are discussed in the context of programs at Statistics Canada. For survey-based estimates, the need for developing an overall strategy is stressed and salient features of survey design that have an impact on small area data are highlighted in the context of redesigning a household survey. A brief review of estimation methods with their strengths and weaknesses is also presented.

KEY WORDS: Sample design strategy; Design estimates; Model estimates.

1. INTRODUCTION

For decades, administrative records and censuses were the main sources of data used for policy and planning for both large and small areas. These are still the richest source of statistical data at small area levels in most countries. During the forties and fifties, however, as the reliance on sample surveys increased, survey based estimates complemented the traditional sources because they provide more timely and cost efficient statistical data in a variety of subject matter fields. Although designed to provide reliable estimates primarily at larger area levels such as national and provincial, increasingly such surveys are being used to meet the growing demands for more timely estimates for various types and sizes of domains. No technical problem arises as long as these domains are large enough (*e.g.*, age-sex groups, larger cities and sub-provincial regions) to yield estimates of acceptable reliability. If data are needed for small domains, however, particularly if such domains cut across design strata, special estimation problems arise and several methods have recently been proposed to deal with such problems.

The main message of this paper is to emphasize the need to look at the problem of small area data in its entirety. Small area needs should be recognized at the early stages of planning for large scale surveys. The sampling design should include special features that enable production of reliable small area data using design or model estimators. The handling of this growing challenge to statistical agencies at the estimation stage should be viewed as a last resort.

In section 2, we discuss data needs and the three main sources of socio-economic data in the Canadian context, namely, the census, administrative records and surveys. Section 3 identifies some technical issues regarding the three sources of data and highlights the problems of quality measures and their interpretation. Then a need for

developing an overall strategy that includes the planning, designing and estimation stages in the survey process is highlighted in section 4. Two aspects of the design, namely, clustering in a multi-stage sample design and sample allocation are discussed. In section 5, we present some sample design options being incorporated during the current redesign of the Canadian Labour Force Survey, the largest monthly household survey conducted by Statistics Canada, with a view to enhancing the survey capacity to provide better quality small area data. The purpose of section 6 is to review the many different approaches to estimation for small areas. We also suggest some new estimators and provide comments on the strengths and weaknesses of various domain estimators. A cautious approach towards the use of model estimators is stressed.

2. INFORMATION NEEDS AND DATA SOURCES

As the country's national statistical agency, Statistics Canada plays an integral role in the functioning of Canadian society. While guaranteeing the confidentiality of individual respondents' data, the agency's information describes the economic and social conditions of the country and its people. Its economic, demographic, social and institutional statistics programs produce reliable data on many aspects of life at the national, provincial, and sub-provincial levels for use by federal and provincial governments, private institutions, academics and the media. With increases in the planning, administration and monitoring of social and fiscal programs at local levels, there has been increasing demand for more and better-quality data at these levels. Three major sources of social, socio-economic and demographic data with emphasis on small area statistics are briefly discussed below.

¹ M.P. Singh, J. Gambino and H.J. Mantel, Statistics Canada, 16th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Census of Population: The quinquennial census of population provides benchmark data and serves as the richest source of information, available every five years, for small areas and for various characteristics/domains/target groups of policy interest such as ethnic minorities, disabled persons, youth and aboriginal peoples.

Administrative Records: Administrative records are an increasingly important source of statistical data. These are extensively used in the demographic field by statistical agencies to produce local area estimates (Schmidt 1952, Verma and Basavarajappa 1987). In certain areas, such as vital statistics, administrative records are the only source of information for production of statistics at various levels of aggregation. In others, the relative merits of administrative records compared to censuses or surveys as data sources in terms of timeliness and quality of data determine the manner and the extent to which these data sources are used. In addition to direct tabulations, administrative records are used in a number of programs as a source of supplementary information for use in improving the quality of survey-based estimates. They are also being used in the construction of sampling frames for conducting surveys. Examples at Statistics Canada include the Business Register and the Address Register of residential dwellings.

Like the census of population, administrative records are very rich in geographical detail, making them a useful source of information for small area statistics. They are available more frequently and, due to recent technological advances, they are becoming a more cost-effective data source. However, administrative data are based on definitions made for programmatic rather than statistical purposes and their content is limited. Details of a Statistics Canada program for integration and development of an administrative records system to produce statistical outputs are given by Brackstone (1987a, 1987b). Experiences in the use of administrative records in other countries are included in the conference proceedings edited by Coombs and Singh (1987).

Household Surveys Program: Household surveys have long been an important source of economic and social statistics at Statistics Canada. Surveys under this program may be placed in three groups, namely, (i) the Labour Force Survey, (ii) Special Surveys and Supplementary Survey Programs and (iii) Longitudinal/Cyclical Surveys. These surveys are briefly introduced below indicating the scope for small area statistics in general.

Starting as a quarterly survey in 1945, the Canadian Labour Force Survey (LFS) became a monthly survey in 1952. The information provided by the survey has expanded considerably over the years and currently it provides a rich and detailed picture of the Canadian labour market. In addition to providing national and provincial estimates the survey regularly releases estimates for subprovincial areas. Regular estimates of standard labour market indicators are also in great demand for small areas such as

Federal Electoral Districts, Census Divisions and Canada Employment Centres. These estimates are used by both federal and provincial governments in monitoring programs and allocating funds and other resources among various political and administrative jurisdictions.

Because of cost considerations, the LFS is heavily used as a vehicle for conducting *ad hoc* and periodic surveys at the national and provincial levels in the form of supplementary or special surveys. In the case of supplements, the LFS respondents themselves are asked additional questions, whereas for special surveys a different set of households is selected using the LFS frame. Both special and supplementary surveys are usually sponsored by other government departments and are conducted on a cost-recovery basis. For these surveys, the demands for small area statistics differ greatly from survey to survey, and generally the demands seem to be less pressing than those from the LFS itself.

Statistics Canada conducts a General Social Survey (GSS) annually to serve, in a modest way, the growing data needs on topics of current social policy interest. The GSS program (Norris and Paton 1991) consists of five survey cycles, each covering a different core topic, repeated every five years. Because of the limited size of sample (10,000 households nationally) the focus of the GSS is on estimates at the national level and on analytical statistics.

Longitudinal/panel surveys are new in the Canadian context. Statistics Canada has started two longitudinal surveys that will enrich the household survey program greatly, namely, the Survey on Labour and Income Dynamics and the National Population Health Survey. Both are large scale panel surveys and they are already creating expectations for data at sub-provincial and local area levels.

3. ISSUES IN DOMAIN ESTIMATION

There are numerous policy and technical issues that need to be addressed in the provision of small area statistics. The seriousness of these issues may vary from agency to agency and from one application to the next within the same agency depending on data quality and release policies. These issues are relevant for national and provincial estimates, but they assume higher significance in the context of small area statistics. As Brackstone (1987a) notes "on the issue of small area data evaluation, it is worth noting that error in small area estimates may be more apparent to users than error in national aggregates... at a local area level, there will be critics quick to point out deficiencies... it is true that for small areas, where estimation is more difficult, scrutiny of estimates is also more intensive". Several research and developmental studies on small area estimation are described in two volumes, one edited by Platek *et al.* (1987), and the other by Platek and Singh (1986). For a

recent overview of small area estimation techniques currently being used in United States federal statistical programs see U.S. Statistical Policy Office (1993).

Use of Administrative Records: Federal and provincial government policies are the prime factors that influence the supply as well as the demand for small area data in most situations. On the supply side, government program driven administrative records contain a wealth of statistical information that can be used to produce local area data. Examples of files being used in the Canadian context are: Family Allowance, Unemployment Insurance, Income Tax, Health, Education, Old Age Security. Income-related statistics are produced at the local area level on a regular basis. Any **change** in government policy and associated programs can have immediate impact, for better or worse, on the coverage, availability, timeliness or quality of statistics derived from the corresponding administrative records. On the demand side, as noted earlier, governments need local area data for planning, implementing and monitoring their policies.

Conceptual issues: Quite frequently, conceptual and definitional issues in a data series are confounded with sampling and estimation problems. For example, consider the Unemployment Insurance (UI) system in Canada. UI regulations stipulate different qualification and requalification periods depending on the unemployment rate in a given region such that regions with higher unemployment rates require shorter qualifying periods of continuous employment. The estimates of regional unemployment rates derived from the LFS are used in determining the eligibility for an individual to receive benefits. These local area estimates are thus continually under close scrutiny by the public and the media. Such scrutiny however refers more often to **conceptual** issues rather than estimation issues per se; aspects such as the treatment in the survey questionnaire of discouraged workers, lay-offs and job search methods are questioned.

Use of Models and Related Quality Measures: Domain estimates are produced for virtually all large scale surveys, and as long as design estimators, *i.e.*, approximately design-unbiased estimators are of acceptable quality, no problem arises. We consider two classes of design estimators. Following Schaible (1992), **direct** estimators refer to estimators which use values of the study variable only for the time period of interest and only from units in the domain (*e.g.*, the regression estimator with slope estimated using only data from the domain). Such estimators may, and often do, use information on one or more auxiliary variables from other domains or other time periods, and are design unbiased or approximately so. The second class of design estimators, **modified** direct estimators, may use information from other domains on both the auxiliary and the study variable but still retain the property of design unbiasedness or approximate unbiasedness (*e.g.*, the regression estimator with slope estimated using the whole

sample). There is a growing literature on **indirect** (or **model**) **estimators**, that is, estimators which use information on both the study and auxiliary variables from outside the domain and/or the time period of interest without any reference to their design unbiasedness properties.

Most producers and users of survey data are accustomed to design estimators and the corresponding design-based inferences. They interpret the data in the context of repeated samples selected using a given probability sampling design, and use estimated design-based **cvs** (coefficients of variation-square root of design variance divided by the design estimate) as the measures of data quality. For situations where either domains are too small or the sampling design did not foresee production of small area estimates, the design estimates may lead to large design **cvs** and model estimates may be the only choice if the survey-based estimates have to be provided for individual domains. A major challenge for statisticians is how to estimate, compare and explain to the users the relative precision of estimates from a survey that produces a large number of estimates at the national, subnational and large and small domain levels, most using design estimators but a few using model estimators. The model-based **cvs** (square root of design variance of model estimate divided by the model estimate) may convey a completely different message and may be several times lower than the corresponding design-based **cvs** for the same small area and in many cases, lower than the design-based **cvs** for much larger areas.

For model estimators, it is usually straightforward to derive expressions for the corresponding mean square errors (*i.e.*, design variance + square of the design bias). Estimation of these expressions, with an adequate degree of reliability, is a different matter. If we follow the argument that the data (*e.g.*, sample size) for such domains are inadequate for producing design estimates, it is unlikely that they would be adequate for producing design estimates of the corresponding variances and biases. As the estimation of bias is relatively more difficult, some authors seek design consistent model estimators, implying perhaps that bias can be ignored. However, if the sample size within the domain is sufficiently large to make the model estimator consistent, then the design estimator itself should give reliable estimates for the domain. For model estimators, suggestions have been made to use estimates of average mean square error computed over all domains. As the need for estimates for different domains usually arises because these domains are thought to be different from each other, a challenging task is to explain why estimates from all such domains are given the same degree of reliability. Another possibility is to construct indirect model-based estimates of the variance and bias of the model estimators for **individual** domains. Finding suitable methods of estimating mean square error for individual domains should be a research priority. Another serious concern for survey practitioners is how to guard against model failures. This

suggests a need for research into model validation for complex survey situations. Further, for model estimators that use data on study variables for periods other than the time period of interest, estimates of **change** over different time periods would be of questionable quality; see Schaible (1992). Also, model estimators that borrow strength from other domains in the larger area will suffer a similar drawback when comparing differences in the two domains within the large area.

Issue of Privacy: In order to construct rich data bases for providing small area statistics, it is sometimes necessary to combine census, survey and/or administrative records. This necessitates linkage of records obtained from different sources. However, given the public's concern about privacy, record linkages should be carried out only after careful examination of all their implications. Under the Statistics Act, Statistics Canada may have access to administrative records of other departments for statistical purposes. But even for statistical purposes, as Fellegi (1987) notes, "we should have rigorous and auditable review procedures to ensure that we only carry out record linkage where the resulting privacy invasion is clearly outweighed by the public good from the new statistical information".

4. NEED FOR AN OVERALL STRATEGY

Even though large scale surveys are designed primarily for national and provincial estimates, it is rare that the estimates from such surveys relate only to the national/provincial populations as a whole. That is, invariably, such surveys are used to produce estimates for various **cross-classified** domains and in some cases for **areal** domains (e.g., subprovincial) as well. In many cases, no special attention is paid to achieving a desired level of precision at the domain level either at the design or the estimation stage as long as the reliability is (believed to be) within reasonable limits. Problems arise when the cross-classified domain refers to a rare subpopulation or when the areal domain refers to a small area in which case either no estimates are possible/available or the estimates are of questionable quality. In a number of cases, this may happen simply because not enough attention was paid to these needs at the start of the survey planning process. If small area data needs are to be served using survey data then there is a need to develop an overall strategy that involves careful attention to meeting these needs at the planning, sample design and estimation stages of the survey process. For discussion of the design and estimation aspects, we will classify domains into the following two types:

Planned domains: In sampling terms these are individual strata or groups of strata for which desired samples have been planned. In the Canadian context these are typically subprovincial regions, such as Economic Regions, Unemployment Insurance Regions, and Health Planning Regions.

In other cases, such domains could be larger counties, districts or similar subprovincial regions.

Unplanned domains: These are areas that were not identified at the time of design and thus may cut across design strata. Such domains can be of any size and they may create special estimation problems.

Planning: As noted earlier, the data demands from continuing periodic surveys such as the LFS are relatively much higher than from *ad hoc* surveys. In the case of periodic surveys that are redesigned every five or ten years, a suitable strategy can be developed during survey redesigns, since, in such cases, statistical agencies are usually in a much better position to project future small area data needs based on past demands. For *ad hoc* surveys, designers should include the establishment of such needs as an integral part of objective setting for the survey. Thus, in both cases, survey designers should establish the desired degree of precision, not only for national and provincial level estimates, but also for the domains of interest.

The first step of a strategy, in terms of the provision of small area data, will depend on the extent to which domains are identified in advance so that they can be treated as planned domains at the time of the design (or redesign) of the survey. If budgetary considerations do not permit reliable estimates for certain very small domains, then the option of either collapsing domains, pooling estimates over different surveys or not providing the estimates at all should be given serious consideration by survey designers in discussions with the survey sponsors. Some domains cannot be determined in advance. These unplanned domains should be handled through special estimation methods.

Sample design: In practice, it is rare that a design is optimal either for the national or provincial levels or for a single subject matter of interest. Usually varying degrees of compromise are introduced at different stages of sampling and the data collection process to satisfy theoretical and operational constraints. Depending on the data needs, estimates for domains should also form an integral part of this compromise. We will discuss two ways of taking small area data needs into account at the design stage, namely, sample allocation and the degree of clustering of the sample.

Allocation Strategy: In general, an optimum allocation strategy for national level estimates allocates samples to provinces approximately in proportion to their population. The reliability of estimates for smaller provinces in such cases suffers. Therefore a compromise allocation is usually preferred. There are different ways in which this compromise can be achieved depending on the emphasis placed on sub-national estimates. Small reductions in sample sizes for larger provinces usually have little effect on the reliability of data for such provinces (or the national level data) but the corresponding sample increase in smaller provinces has significant impact on the reliability of their data.

The same principle holds for planned domains within the provinces. This is because optimum allocations in most situations are flat and the designers can exploit this feature by reallocating sample from the larger areas to planned domains that are smaller in size.

Clustering: Large scale household surveys usually involve stratified multistage designs with relatively large primary sampling units in order to make the design cost-efficient for national and provincial statistics. Such designs are thus highly clustered and, therefore, detrimental to the production of statistics for unplanned areal domains in the sense that, due to chance, some domains may be sample-rich while others may have no sample at all. Given the importance of domain estimates, attempts should be made to minimize the clustering in the sample. The following factors are important in this context: choice of frame, choice of sampling units and their sizes, number and size of strata and stages of sampling. The goal should be to make the design effects as low as possible given the operational constraints.

Estimation: No matter how much attention is paid to domain estimates at the early stages of planning and designing a particular survey, there will always be some smaller domains for which special estimation methods will be required for producing adequate estimates. Recently, synthetic estimators, which borrow strength from domains that resemble the domain of interest, have attracted a good deal of attention. However, since synthetic estimators are very sensitive to the assumption that domains resemble each other, even a small departure from the assumption can make the design bias high and put their use in question. Probability samplers, conscious of design bias, have suggested combinations of direct and synthetic estimators, with a view to addressing the design bias problem while trying to retain the strengths of the synthetic estimator. Empirical Bayes and similar techniques have been used to assign a weight to each component in the combined estimators. A brief review of these developments is given in section 6 on estimation.

5. SAMPLE DESIGN CONSIDERATIONS

5.1 Introduction

The small area problem is usually thought of as one to be dealt with via estimation. However, as was noted in the previous section, there are opportunities to be exploited at the survey design stage. This section uses the Canadian Labour Force Survey (LFS) to illustrate this.

The current LFS design: The Canadian Labour Force Survey is a monthly survey of 59,000 households which are selected in several stages using various methods. The ultimate sampling unit, the household, remains in the sample for six months once it is selected and is then

replaced. Higher stage units (primary sampling units (PSU), clusters) also rotate periodically. Each of Canada's ten provinces is divided into economic regions (ER) which the LFS further divides into self-representing areas (medium and large cities) and non-self-representing areas (the rest of the ER). Stratification and sample selection take place within these areas, and the number of stages of sampling as well as the units of sampling differ between these two types of area. For example, in areas outside cities, there are three stages of sampling, whereas there are only two in the cities. For a detailed description of the current LFS design, refer to Singh *et al.* (1990).

5.2 Sampling Stages and Sampling Units

Area frames are usually associated with clustered sampling, *i.e.*, the first-stage units of selection are typically land areas containing a number of second-stage units. If a list of the second-stage units becomes available, then sampling directly from the list becomes possible, leading to a less clustered sample. This will result not only in improved estimates (due to lower design effects) but also in better small area estimates for unplanned domains. The latter holds since, by spreading the sample more evenly, it is more likely that an unplanned areal domain will contain some selected units. In contrast, in a clustered design we are often faced with a situation where one domain has sufficient sample because it happens to contain sampled clusters while a similar domain happens to have too few or no sampled clusters to produce good estimates.

To reduce clustering in the LFS we investigated two options: (i) the possibility of replacing the area frame (with its two stage design) in the larger cities with a list frame using the Address Register and (ii) reducing the sampling stages in rural areas and smaller urban centres. The Address Register, created to improve the coverage of the 1991 Canadian census (Swain, Drew, Lafrance and Lance 1992), consists of a list of addresses, telephone numbers and geographical information for dwellings by census enumeration area (EA). One option involved the selection of a stratified simple random sample of dwellings from the Address Register frame. This sample could then be supplemented with a sample selected from a growth frame which comprises a set of dwellings that are not in the post-censal address register. Handling of growth became the major stumbling block in pursuing option (i) as no cost-efficient method could be devised and tested in time for the current redesign. However, an updating strategy for the post-censal Address Register is still being investigated for future censuses and surveys.

With regard to option (ii), in keeping with the idea that less clustering is better for small area estimates, changes in the units and reduction in the stages of sampling were investigated for the areas outside the cities. Due to the changes that have taken place in data collection techniques,

namely, from face-to-face interviewing to telephone and computer assisted interviewing, the cost-variance analyses from the past are no longer relevant. More than 80 percent of LFS interviews are now conducted by telephone. With the increase in telephone interviewing and the resulting decrease in travel, it became feasible in almost all cases to eliminate the current PSU stage and to sample EAs directly.

5.3 Stratification

One approach to stratification, similar in spirit to the above discussion on PSU size, is to replace large strata by many small ones. The hope is that a redefined domain or an unplanned domain will contain mostly complete strata. This will make the sample size in the domain more stable.

There may be several overlapping areas for which estimates are required. For example, each Canadian province is partitioned into both Economic Regions (ER) and Unemployment Insurance regions (UIR). One way to deal with this situation is to treat all the areas created by the intersections of the partitions as strata. In the Canadian case, for example, the 71 ERs and 61 UIRs yield 133 intersections, a manageable number. In some cases, however, the number of intersections may be too large to handle effectively. In addition, some of the intersections may have very small populations, making them unusable as strata.

By combining decreased clustering with smaller strata, we hope to have a design which is better able to meet small area needs. For example, the design should provide more flexibility in satisfying both ER and UIR requirements efficiently and in dealing with future changes in the definition of regions.

5.4 Allocation

If the definitions of small areas are known in advance, we may be able to treat them as planned domains and take them into account when designing the survey. The survey designer may endeavour to allocate sufficient sample in each small area to make the production of reliable estimates feasible. For large surveys such as the Canadian Labour Force Survey, this approach can, at least in theory, make the production of a great many small area estimates feasible. With a monthly sample of 59,000 households, and assuming that, say, 100 households per month are needed to produce reliable quarterly estimates, the country can be divided into about 600 non-overlapping areas, each guaranteed to have sufficient sample. Unions of such areas will also have enough sample to produce reliable monthly estimates.

Various sample allocation strategies are possible. In a top-down approach, once a provincial sample size is determined, the sample is allocated among the sub-provincial regions. However, it may turn out that it is not possible to satisfy the requirements for the reliability of sub-provincial

estimates for the given provincial sample size. In a bottom-up strategy, the sample would be allocated to sub-provincial regions first in such a way that reliability objectives for each region are satisfied. As a result, we would expect comparable sample sizes in each sub-provincial region. This approach may result in a provincial sample size that is bigger than the one specified in the top-down approach. Regardless of which of the two strategies is used, adjustments to the initial allocations will usually be required. The resulting allocation will likely resemble a compromise between proportional allocation and equal allocation. In practice, the survey designer must perform a complex juggling act among provincial reliability requirements, sub-provincial requirements for one or more sets of regions, total survey costs and in-the-field details.

The approach taken in the current LFS redesign may be useful in other surveys as well. The sample was allocated in two steps: first, a core sample of 42,000 households was allocated to produce good estimates at the national and provincial levels; then the remaining sample was allocated to produce the best possible sub-provincial estimates. The resulting compromise allocation will produce reliable estimates for almost all planned domains. The compromise resulted in only minor losses at the provincial level and substantial gains at the subprovincial level. For example, the expected CVs for 'unemployed' for Ontario and Quebec are 3.2 and 3.0 per cent, respectively, instead of 2.8 and 2.6. The corresponding figures for Canada are 1.51 and 1.36. Optimizing for the provincial level yields CVs as high as 17.7 per cent for UI regions. With the compromise allocation, the worst case is 9.4 per cent.

Sample redistribution: There is usually some scope for moving sample from one area to another. For example, reducing the sample size by 1,000 households in a large province and making a corresponding increase in a small province will cause a marginal deterioration in the quality of provincial estimates in the former but will improve the estimates in the latter significantly. Similar movements of sample can be attempted within province.

5.5 Other Considerations

Change in definitions of small areas: Survey designers are faced with the fact that the definitions of planned domains may change during the life of a design and they may then have to treat the new domains as unplanned domains. For example, it is quite possible that the definitions of Unemployment Insurance Regions will change two or three years after the new LFS design is introduced in 1995. To deal with this at the design stage, the best that the survey designer can do is to choose as building blocks areas which are standard (e.g., census-defined areas whose definitions are fairly stable) and hope that the redefined regions are unions of these standard areas. This is the approach that was taken in the current LFS redesign.

An alternative is to adopt an update strategy. This entails a reselection of units, doing it in such a way that the overlap between the originally selected units and the newly selected ones is maximized. By taking this approach, the number of new units that have to be listed is minimized. This also minimizes other field disruptions such as the need to hire new interviewers.

6. ESTIMATION

The purpose of this section is to review some of the different approaches to estimation of totals for small areas. No attempt is made to provide an exhaustive review; the discussion indicates the trend of developments in small area estimation research. For a detailed review, see the recent paper by Ghosh and Rao (1993). To facilitate this review we will classify small area estimation methods into two types. This is just one of many possible classification schemes. The first class of estimators we call design estimators, *i.e.*, (approximately) design unbiased estimators, which includes direct and modified direct estimators. As noted earlier, design estimators are often unsatisfactory, having a large variance due to small sample sizes (or even no sample at all) in the small areas. The second class we call indirect (or model) estimators, and it includes synthetic and combined estimators. Some of these estimators are compared empirically in an earlier version of this paper by Singh, Gambino and Mantel (1992).

6.1 Design Estimators

Direct Estimators: Direct small area estimators are based on survey data from only the small area, perhaps making use of some auxiliary data from census or administrative sources in addition to the survey data. The simplest direct estimator of a total is the expansion estimator,

$$\hat{Y}_{e,a} = \sum_{i \in s_a} w_i y_i, \quad (6.1)$$

where s_a is the part of the sample in small area a and w_i is the survey weight for unit i . This estimator is unbiased; however, it may have high variability due to the random sample size in area a .

If the population size N_a is known then a post-stratified estimator,

$$\hat{Y}_{pst,a} = N_a \sum_{i \in s_a} w_i y_i \Big/ \sum_{i \in s_a} w_i = N_a \hat{Y}_{e,a} / \hat{N}_{e,a} = N_a \bar{y}_{e,a}, \quad (6.2)$$

may be used. This estimator is more stable than the expansion estimator; however, there may be some ratio estimation bias in complex surveys.

If the sampling scheme is stratified and the $N_{h,a}$ are known, where $N_{h,a}$ is the population size in stratum h and small area a , an alternative post-stratified estimator is $\hat{Y}_{st,pst,a} = \sum_h (N_{h,a} \sum_{i \in s_{h,a}} w_i y_i / \sum_{i \in s_{h,a}} w_i) = \sum_h N_{h,a} \hat{Y}_{h,e,a} / \hat{N}_{h,e,a} = \sum_h N_{h,a} \bar{y}_{h,a}$. The strata may also be post-strata instead of design strata.

Ratio estimation is similar to post-stratified estimation, the difference being that another auxiliary variable is used in place of the population counts N_a and $N_{h,a}$. For example, if x is a covariate for which the small area totals, X_a , or the stratum small area totals, $X_{h,a}$, are known then we may define the ratio estimators

$$\hat{Y}_{r,a} = X_a \hat{R}_a \quad \text{and} \quad \hat{Y}_{st,r,a} = \sum_h X_{h,a} \hat{R}_{h,a}, \quad (6.3)$$

where $\hat{R}_a = \hat{Y}_{e,a} / \hat{X}_{e,a}$ is an estimate of the ratio Y_a / X_a and $\hat{R}_{h,a} = \hat{Y}_{h,e,a} / \hat{X}_{h,e,a}$.

A regression estimator attempts to account for differences between small area subpopulation and subsample values of the covariates via an estimated regression relationship between the variate of interest, y , and the covariates, x . An advantage of regression type estimation is that it is easily extended to vector covariates. The estimator is given by

$$\hat{Y}_{reg,a} = \hat{Y}_a + \hat{\beta}_a (X_a - \hat{X}_a), \quad (6.4)$$

where \hat{Y}_a may be an expansion or post-stratified estimator, \hat{X}_a must be calculated in the same way as \hat{Y}_a , and $\hat{\beta}_a = \sum_{i \in s_a} v_i^{-1} w_i y_i x_i' \{ \sum_{i \in s_a} v_i^{-1} w_i x_i x_i' \}^{-1}$ where v_i are given weights for the regression. Note that $\hat{\beta}_a = \hat{R}_a$ when x is scalar and $v_i = x_i$. When \hat{Y}_a and \hat{X}_a are expansion estimators this estimator is also called the generalized regression estimator. Approximate design unbiasedness of this estimator follows from that of \hat{Y}_a and \hat{X}_a .

As with the ratio type estimators, regression type estimation may also be applied within design strata or post-strata.

Modified Direct Estimators: Modified direct estimators may use survey data from outside the domain; however, they remain approximately design unbiased. By a modified direct estimator we mean a direct estimator with a synthetic adjustment for model bias; since the adjustment would have approximately zero expectation with respect to the design, the modified estimator is approximately design unbiased if the direct estimator is. An example is obtained by replacing $\hat{\beta}_a$ in (6.4) by a synthetic estimator $\hat{\beta} = \sum_{i \in S} v_i^{-1} w_i y_i x_i' \{ \sum_{i \in S} v_i^{-1} w_i x_i x_i' \}^{-1}$; we will denote this estimator by $\hat{Y}_{sreg,a}$. $\hat{\beta}$ would generally be more stable than $\hat{\beta}_a$; the choice between them would depend on the size of the variance of $\hat{\beta}_a$ relative to the variation in the β_a s over areas a . A compromise is to take a weighted average $\lambda_a \hat{\beta}_a + (1 - \lambda_a) \hat{\beta}$ where λ_a is suitably chosen;

options for the choice of λ_a are discussed under combined estimators in Section 6.2. A second example is obtained by replacing $\hat{\beta}_a$ in (6.4) by $\hat{R} = \hat{Y}_e / \hat{X}_e$; note that \hat{R} is a special case of $\hat{\beta}$ where x is scalar and $v_i = x_i$.

6.2 Indirect Estimators

Synthetic Estimators: Synthetic estimation methods are based on an assumption that the small area is similar in some sense to another area, often a larger area which contains it. Estimates for the other area would generally be more reliable than those for the small area. The resulting synthetic estimator would then have small variance, though it may be badly biased if the underlying assumption is violated.

One of the simplest synthetic estimators arises from the assumption that the small area mean is equal to the overall mean. This leads to the mean synthetic estimator

$$\hat{Y}_{syn,m,a} = N_a \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i = N_a \bar{y}. \quad (6.5)$$

A more common synthetic estimator is based on stratification or post-stratification,

$$\hat{Y}_{syn,st,m,a} = \sum_h N_{h,a} \sum_{i \in s_h} w_i y_i / \sum_{i \in s_h} w_i = \sum_h N_{h,a} \bar{y}_h.$$

As with direct estimators, ratio synthetic estimation may be based on other auxiliary data besides the population counts N_a or $N_{h,a}$. For example, the common ratio synthetic estimators based on a covariate x are defined as

$$\hat{Y}_{syn,r,a} = X_a \hat{Y}_e / \hat{X}_e \quad \text{and} \quad \hat{Y}_{syn,st,r,a} = \sum_h X_{h,a} \hat{Y}_{h,e} / \hat{X}_{h,e}, \quad (6.6)$$

where $\hat{Y}_e = \sum_{i \in s} w_i y_i$ is the expansion estimator of the population total for y and $\hat{Y}_{h,e} = \sum_{i \in s_h} w_i y_i$. \hat{X}_e and $\hat{X}_{h,e}$ are similarly defined. These estimators have been studied by Gonzalez (1973), Gonzalez and Waksberg (1973) and Ghangurde and Singh (1977, 1978), among others.

Singh and Tessier (1976) suggested an alternative ratio synthetic estimator, using X instead of \hat{X}_e , defined as

$$\tilde{Y}_{syn,r,a} = X_a \hat{Y}_e / X. \quad (6.7)$$

Both $\hat{Y}_{syn,r,a}$ and $\tilde{Y}_{syn,r,a}$ have the same synthetic bias and the ratio bias in $\tilde{Y}_{syn,r,a}$ will be negligible for large samples. The choice between these two estimators depends on ρ , the correlation of \hat{Y}_e and \hat{X}_e . It can be shown that for large samples $V(\tilde{Y}_{syn,r,a}) \leq V(\hat{Y}_{syn,r,a})$ if $\rho \geq 0.5c_x/c_y$, where c_x and c_y are the coefficients of variation of \hat{X}_e and \hat{Y}_e , respectively. In most cases, when ρ is high or the population is skewed, $\tilde{Y}_{syn,r,a}$ would be preferred; however, when c_x is high and the correlation is only moderate, $\hat{Y}_{syn,r,a}$ may be the better choice.

In some situations information on a second auxiliary variable (z) in addition to x may be available. Then a bivariate ratio synthetic estimator may be constructed:

$$\hat{Y}_{syn,r,a}^{(2)} = \gamma_a X_a \hat{Y}_e / \hat{X}_e + (1 - \gamma_a) Z_a \hat{Y}_e / \hat{Z}_e, \quad (6.8)$$

where γ_a is suitably chosen. Extensions to a multivariate ratio synthetic estimator may be considered following Olkin (1958).

Regression synthetic estimation is similar to ratio synthetic,

$$\hat{Y}_{syn,reg,a} = \hat{\beta} X_a,$$

$$\hat{\beta} = \sum_{i \in s} v_i^{-1} w_i y_i x_i' \left\{ \sum_{i \in s} v_i^{-1} w_i x_i x_i' \right\}^{-1}. \quad (6.9)$$

Again, regression synthetic estimation may also be applied within design strata or post-strata. Royall (1979) suggested a slight variation, $\hat{Y}_{syn,Roy,a} = \sum_{i \in s_a} y_i + \hat{\beta}(X_a - \sum_{i \in s_a} x_i)$, where the sum of y -values for only units not included in the sample is estimated synthetically.

Remark: The examples of modified direct estimators presented in Section 6.1 can also be considered to be ratio or regression synthetic estimators with a design-based adjustment to correct for bias. For example, we may write $\hat{Y}_{sreg,a} = \hat{Y}_{syn,reg,a} + (\hat{Y}_a - \hat{\beta} \hat{X}_a)$ where $\hat{Y}_a - \hat{\beta} \hat{X}_a$ is an estimate of the bias of $\hat{Y}_{syn,reg,a}$. Similarly, $\hat{Y}_{sreg,a}$ can also be written as the Royall estimator, $\hat{Y}_{syn,Roy,a}$, with a design-based adjustment for bias.

Purcell and Kish (1980) discuss another type of synthetic estimation which they call SPREE (structure preserving estimation) for small area estimation of frequency data. Detailed historical counts, perhaps from a census, are combined with less detailed current survey estimates to produce detailed estimates of current counts. The assumption here is that certain relationships among the detailed counts are stable over time.

Combined Estimators: By a combined estimator we mean a weighted average of a design estimator and a synthetic estimator,

$$\hat{Y}_{com,a} = \lambda_a \hat{Y}_{des,a} + (1 - \lambda_a) \hat{Y}_{syn,a}, \quad (6.10)$$

where λ_a is suitably chosen. The aim here is to balance the potential bias of the synthetic estimator against the instability of the design estimator. There are three broad approaches which may be used to define the weights λ_a in (6.10); they may be fixed in advance, sample size dependent, or data dependent.

The first and simplest approach to weighting is to fix the weights in advance, for example, to take a simple average. However, this does not make any allowance for

the actual observed reliability of the design estimator. For some realized samples the design estimator for small area a is more reliable than for other realized samples. The weight given to the design estimator should reflect this.

The second general approach to weighting of the design and synthetic parts is called sample size dependent, in which the weights are functions of the ratio $\hat{N}_{e,a}/N_a$. Another possibility, not considered here, is to base the weights on the realized sample values of a covariate x ; for example, the weight could be a function of $\hat{X}_{des,a}/X_a$ or of $S_{x,a}^2/\sigma_{x,a}^2$ where $S_{x,a}^2$ is the realized variance of $\hat{X}_{des,a}$, conditional on $\hat{N}_{e,a}$ or some other relevant aspect of the realized sample, and $\sigma_{x,a}^2$ is the unconditional variance of $\hat{X}_{des,a}$.

Some specific estimators in this class have been proposed earlier. Drew, Singh, and Choudhry (1982) proposed the sample size dependent estimator

$$\hat{Y}_{ssd,r,a} = \lambda_a \hat{Y}_{r,a} + (1 - \lambda_a) \hat{Y}_{syn,r,a}, \quad (6.11a)$$

where

$$\lambda_a = \begin{cases} 1 & \text{if } \hat{N}_{e,a} \geq \delta N_a \\ \hat{N}_{e,a}/\delta N_a & \text{otherwise} \end{cases} \quad (6.11b)$$

and δ is subjectively chosen to control the contribution of the synthetic component. Särndal (1984) suggested

$$\hat{Y}_{ssd,reg,a} = \lambda_a \hat{Y}_{sreg,a} + (1 - \lambda_a) \hat{Y}_{syn,reg,a}, \quad (6.12)$$

where $\lambda_a = \hat{N}_{e,a}/N_a$. Rao (1986) suggested a modification to this in which λ_a would be taken to be 1 whenever $\hat{N}_{e,a} \geq N_a$. Särndal and Hidioglou (1989) refined Rao's suggestion by taking $\lambda_a = (\hat{N}_{e,a}/N_a)^{h-1}$ when $\hat{N}_{e,a} < N_a$, where h is chosen judgementslly to control the contribution of the synthetic component.

It is the bias of the synthetic component that is of concern when using these sample size dependent estimators in practice. The weight associated with the synthetic component should be such that the bias is kept within reasonable limits. For example, the sample size dependent estimator of Drew, Singh and Choudhry (1982), with generalized regression estimation replacing the ratio estimation and $\delta = 2/3$, is currently used in the Canadian Labour Force Survey to produce domain estimates. For a majority of domains the weight attached to the synthetic component is zero as the direct estimator itself provides the required degree of reliability. For other domains the weight attached to the synthetic component is about 10% on average and never exceeds 20%. Depending on the risk of bias that one is willing to take, δ may lie in the range $[2/3, 3/2]$ for most practical situations.

The third approach to weighting we call data dependent. The optimal weights for combining two estimators generally depend on the mean squared errors of the estimators and

their covariance. These quantities would generally be unknown but may be estimated from the data. For our combined estimators this would usually require some modelling of the bias of the synthetic part. An early and well known example of this approach is due to Fay and Herriot (1979). They model the biases of the synthetic estimators for the small areas as independent random effects with an unknown but fixed variance. To be more specific, if $\hat{Y}_{des,a}$ is the design estimator then they consider the model $Y_a = X_a \beta + \alpha_a$ and $\hat{Y}_{des,a} = Y_a + \epsilon_a$ where $\alpha_a \sim (0, \sigma^2)$, $\epsilon_a \sim (0, \nu_a^2)$, and α_a and ϵ_a are independent and uncorrelated over a , σ^2 is unknown and ν_a^2 are assumed known (in practice they would need to be estimated). For a given value of σ^2 the optimal weights for combining $\hat{Y}_{des,a}$ and $X_a \hat{\beta}$ can be calculated. An estimate of σ^2 is obtained by the method of fitting constants and substituted into the optimal weights. Some protection against model mis-specification is obtained by truncating the resulting estimate if it deviates from the direct estimate by more than a specified multiple of ν_a . Schaible (1979) and Battese and Fuller (1981) also consider empirically estimated optimal weights λ_a in (6.12) based on similar random effects models for the small area totals.

Prasad and Rao (1990) provide an estimator of the mean square error of the Fay-Herriot estimator which makes allowance for the estimation of the variance components. Kott (1989) proposes a design consistent estimator of the mean square error, but finds it to be very unstable.

Another alternative is to use historical data to calculate the weights; this has the advantage that the weights may be more stable than if they are estimated from current survey data; however, there is an underlying assumption that the optimal weights are stable over time.

Remark: In sample size dependent estimation the weights are allowed to depend on the observed size of the subsample s_a , but not on the values of the variate of interest. This non-dependence of the weights on the variate of interest has advantages and disadvantages. An advantage is that the same weights would be used for estimation of totals for all variates of interest; they need to be calculated only once. More importantly, the estimate of the sum of two variables is the sum of the estimates of the two variables. A disadvantage is that the weights do not directly take account of either the reliability of the design estimator for the variate of interest or the likely magnitude of the bias of the synthetic estimator.

Combining data over time: For repeated surveys pooling of data over survey occasions to increase the reliability of estimates is a common practice. Depending on the rotation pattern used for such surveys, significant gains in reliability can be achieved. This pooling or averaging over time is thus of particular interest in the context of domain estimation where reliability is usually low. For domain

estimation in the Canadian Labour Force Survey it is normal practice to use a sample size dependent estimator based on three month average estimates of employed and unemployed. Due to the six month rotation scheme used, as noted earlier, averaging over three months increases the sample size by one third. If samples completely overlap between periods then averaging does not result in any gain in efficiency. For other rotation patterns the sample size for domain estimates could be more than doubled through this process. There is, however, a conceptual problem with pooled estimates, in that such estimates refer to an average of the parameter of interest (*e.g.*, unemployment) over a period of, say, three months.

In composite estimation the current design estimator is combined with the composite estimator for the previous period, updated by an estimate of change based on the common sample. This idea was used, though not in the context of small area estimation, by Jessen (1942), and Patterson (1950), among others. Binder and Hidirolou (1988) provide a review. The weights for the combination are typically estimates of the optimal weights under the assumption that these weights are time stationary. These data dependent weights have the disadvantage that they lead to inconsistency of estimates for different characteristics and their sums.

A recent development in small area estimation techniques is the use of time series methods for periodic surveys. The relationship between parameters of interest for different time periods is modelled and this model is exploited to improve the efficiency of the estimates for the current occasion. In most cases some allowance must also be made, through modelling or otherwise, for the non-independence of samples for different survey occasions due to the sample rotation scheme. Some references for this time series approach are Choudhry and Rao (1989), Pfeffermann and Burck (1990), Singh, Mantel and Thomas (1994) and Singh and Mantel (1991). All of these are generalizations of the Fay-Herriot model which allow the regression parameters, small area effects, and survey errors to evolve over time according to various time series models. The vector of small area estimates that results from this approach can be written as a weighted average of the vector of design estimates and a vector of synthetic estimates which are based on past data and the current values of covariates; however, the matrix of weights would not generally be diagonal so that the estimator for any single small area would generally depend also on the design estimates and synthetic estimates for other small areas.

7. CONCLUSION

To produce adequate survey-based domain estimates that are timely and up to date, sample designers must face several challenging tasks. The first is to convince the

sponsors/program managers that some small area data needs cannot be met as a by-product of a system designed optimally for national/sub-national estimates. Significant gains, which may vary from survey to survey, can be achieved at the domain level at a marginal reduction in reliability at higher levels. There is a need to develop an overall strategy that incorporates desired reliability for the planned domains as well as for higher levels through compromise allocations, and reduced clustering to help improve estimates for unplanned domains. It should be noted that many of the planned domains at design time may become unplanned (revised) over time in the context of continuous surveys.

The overall strategy should also include consideration of both design estimators for larger domains and model estimators for small domains. A model estimator should be preferred over a design estimator only if its mean square error (design variance + bias²) is estimable and it is sufficiently smaller than the corresponding variance of the design estimator. We should have estimates of mean square error for each of the individual domains. An option that statistical agencies can exercise is to pool similar domains or pool estimates over different time periods for the same domain. They may even suppress estimates for some domains on account of data reliability or privacy concerns.

The second challenging task for statisticians is to explain to users the different types of measures of reliability for different sets of estimates from the same survey. It is hoped that with more research on model validation and better estimates of mean square errors, designers will get more confidence in using model estimators for small domains. In the meantime model estimators should be used with caution even if they have significantly smaller coefficients of variation.

Censuses, supplemented by data from administrative records, are likely to remain the primary source of small area socio-economic data, especially for countries having a quinquennial census of population and housing. Also, concerns about problems with conceptual issues in the context of data for administrative records are likely to continue until statistical agencies are given an opportunity to influence the development of the forms used to collect such data. Until then, this immensely rich data source cannot be fully exploited for statistical purposes and more so for domain estimation.

ACKNOWLEDGMENT

We are grateful to Jon Rao for handling this paper as an Associate Editor and to the referees for many constructive suggestions.

REFERENCES

- BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.
- BINDER, D., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Eds. P.R. Krishnaiah and C.R. Rao). New York: Elsevier Science, 187-211.
- BRACKSTONE, G.J. (1987a). Small area data: Policy issues and technical challenges. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley and Sons, 3-20.
- BRACKSTONE, G.J. (1987b). Statistical uses of administrative data: Issues and challenges. *Proceedings: Symposium on Statistical Uses of Administrative Data*, (Eds. J.W. Coombs and M.P. Singh), Statistics Canada, 5-16.
- CHOUDHRY, G.H., and RAO, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data. *Proceedings: Symposium on Analysis of Data in Time*, (Eds. A.C. Singh and P. Whitridge), Statistics Canada, 67-74.
- COOMBS, J.W., and SINGH, M.P. (Eds.) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.
- DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FELLEGI, I.P. (1987). Opening Remarks. *Proceedings: Symposium on Statistical Uses of Administrative Data*, (Eds. J.W. Coombs and M.P. Singh), Statistics Canada, 1-2.
- GHOSH, M., and RAO, J.N.K. (1993). Small area estimation: An appraisal. To appear in *Statistical Science*.
- GHANGURDE, P.D., and SINGH, M.P. (1977). Synthetic estimates in periodic household surveys. *Survey Methodology*, 3, 152-181.
- GHANGURDE, P.D., and SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.
- GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- GONZALEZ, M.E., and WAKSBERG, J. (1973). Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.
- JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin*, 304, 54-59.
- NORRIS, D., and PATON, D. (1991). Canada's General Social Survey: Five years of experience. *Survey Methodology*, 17, 227-240.
- OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. Invited Presentations. New York: Wiley.
- PLATEK, R., and SINGH, M.P. (1986). Small Area Statistics, An International Symposium' 85 (contributed papers), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, University of Ottawa, Canada.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or Domains). *International Statistical Review*, 48, 3-18.
- RAO, J.N.K. (1986). Synthetic estimates, SPREE and best model based predictors. *Proceedings of the Conference on Survey Research Methodology in Agriculture*, American Statistical Association and National Agricultural Statistics Service, USDA, 1-6.
- ROYALL, R.M. (1979). Prediction models in small area estimation. In *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare.
- SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area*, NIDA Research Monograph Series 24, U.S. Department of Health, Education and Welfare, Library of Congress catalogue number 79-600067, 36-53.
- SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos and R. Platek). Warsaw: Central Statistical Office, 95-114.
- SCHMIDT, R.C. (1952). Short-cut methods for estimating county populations. *Journal of the American Statistical Association*, 47, 232-238.
- SINGH, A.C., and MANTEL, H.J. (1991). State space composite estimation for small areas. *Proceedings: Symposium 91, Spatial Issues in Statistics*, Statistics Canada, 17-25.

- SINGH, A.C., MANTEL, H.J., and THOMAS, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 33-43.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.
- SINGH, M.P., GAMBINO, J.G., and MANTEL, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos and R. Platek). Warsaw: Central Statistical Office, 37-75.
- SINGH, M.P., and TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.
- SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register for coverage improvement in the 1991 Canadian Census. *Survey Methodology*, 18, 127-142.
- U.S. STATISTICAL POLICY OFFICE (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21. Prepared by the subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology.
- VERMA, R.B.P., and BASAVARAJAPPA, J.G. (1987). Recent developments in the regression method for estimation of population for small areas in Canada. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley and Sons, 46-61.

COMMENT

W.A. FULLER¹

The authors are to be congratulated on an excellent description of the design and estimation considerations associated with domains. The authors discuss estimation for planned domains, particularly situations in which domain membership can be identified in the frame, and estimation for unplanned domains including domains for which the domain membership cannot be determined from the frame. This is a fine contribution to the growing literature on domain estimation.

The authors give a particularly good description of the planning, data collection, and processing activities associated with surveys conducted by Statistics Canada. Included are the traditional design problems of balancing needs for domain estimation with desire for efficiency at higher levels, the importance of confidentiality in using administrative records in constructing domain estimates, and the importance of definitional compatibility in attempting to combine information from different sources.

The importance of considering domain estimation at the design stage is very well taken and is a point often ignored by authors concentrating on small area estimation. As the authors emphasize, careful design can often enable one to construct estimates for domains in a direct and design consistent manner. I am sure that those actually designing surveys have considered the importance of clustering when designing surveys that will be used for domain estimation, but it is pleasant to see an explicit discussion.

The authors describe several types of estimators for domains. Their classification emphasizes the number of alternatives available to the practitioner. It is possible to use the theoretical mean square errors to provide information on the relative merits of the estimators. As an example of such a comparison, assume a simple random sample of size n selected from a population divided into K domains. Assume that the domain sizes and the domain means of an auxiliary variable, X , are available. Consider the three regression estimators of the domain mean,

$$\hat{\mu}_{(1)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b_i,$$

$$\hat{\mu}_{(2)yi} = \bar{y}_{i.} + (\mu_{xi} - \bar{x}_{i.})b.$$

and

$$\hat{\mu}_{(3)yi} = \bar{y}_{..} + (\mu_{xi} - \bar{x}_{..})b.,$$

where

$$(\bar{x}_{..}, \bar{y}_{..}) = \sum_{i=1}^k N^{-1} N_i (\bar{x}_{i.}, \bar{y}_{i.}),$$

$$(\bar{x}_{i.}, \bar{y}_{i.}) = n_i^{-1} \sum_{j=1}^{n_i} (X_{ij}, Y_{ij}),$$

$$b_i = \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})^2 \right]^{-1} \times \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

$$b. = \left[\sum_{i=1}^k N^{-1} N_i n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})^2 \right]^{-1} \times \sum_{i=1}^k N^{-1} N_i n_i^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{x}_{i.})(Y_{ij} - \bar{y}_{i.}),$$

n_i is the number of observations in domain i , N_i is the population size of domain i , μ_{xi} is the population mean of X for domain i , and $\mu_{x.}$ is the grand population mean of X . In the authors' terminology, the first estimator is a direct regression estimator, the second is a modified direct estimator, and the third is a synthetic estimator. We have

$$\text{MSE}\{\hat{\mu}_{(1)yi} | n_i\} = n_i^{-1} (1 + n_i^{-1}) V\{Y_{\ell j} - \beta_{\ell} X_{\ell j} | \ell = i\} + O(n_i^{-2}),$$

$$\text{MSE}\{\hat{\mu}_{(2)yi} | n_i\} = n_i^{-1} (1 + n^{-1}) V\{Y_{\ell j} - \beta X_{\ell j} | \ell = i\} + O(n^{-2}),$$

$$\begin{aligned} \text{MSE}\{\hat{\mu}_{(3)yi} | n_i\} &= (1 + n^{-1}) \\ &\times \sum_{i=1}^k N^{-2} N_i^2 n_i^{-1} V\{Y_{\ell j} - \beta X_{\ell j} | \ell = i\} \\ &+ (\mu_{xi} - \mu_{x.})^2 V\{b.\} \\ &+ [\mu_{yi} - \mu_{y.} - \beta(\mu_{xi} - \mu_{x.})]^2 + O(n^{-2}), \end{aligned}$$

where $V\{b.\} = E\{(b. - \beta)^2\}$, $V\{a_{\ell} | \ell = i\}$ is the variance of the variable a for domain i ,

$$\beta_i = [V\{X_{\ell j} | \ell = i\}]^{-1} C\{Y_{\ell j}, X_{\ell j} | \ell = i\}$$

¹ W.A. Fuller, Distinguished Professor, Statistical Laboratory and Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa.

and

$$\beta = \left[\sum_{i=1}^k N^{-1} N_i V\{X_{\ell j} | \ell = i\} \right]^{-1} \\ \times \sum_{i=1}^k N^{-1} N_i C\{Y_{\ell j}, X_{\ell j} | \ell = i\}.$$

The estimator $\hat{\mu}_{(1)yi}$ uses only information in the sample of n_i observations. Hence, all properties of the estimator are functions of n_i and of the domain parameters. The regression bias is order n_i^{-1} and the variance is order n_i^{-1} . The estimator $\hat{\mu}_{(2)yi}$ uses the domain means, but the entire sample to estimate the regression coefficient. Hence, the basic variance remains order n_i^{-1} and will be larger than the basic variance of $\hat{\mu}_{(1)yi}$ in those situations where $\beta_i \neq \beta$. However, the second order contribution to the variance is order $n_i^{-1} n^{-1}$ for $\hat{\mu}_{(2)yi}$ and is order n_i^{-2} for $\hat{\mu}_{(1)yi}$. Also, the regression bias for $\hat{\mu}_{(2)yi}$ is order n^{-1} . If the domains were strata, $\hat{\mu}_{(1)yi}$ might be called the separate regression estimator and $\hat{\mu}_{(2)yi}$ might be called the combined regression estimator.

The estimator $\hat{\mu}_{(3)yi}$ is a synthetic estimator and has a variance of order n^{-1} instead of the order n_i^{-1} variance of the first two estimators. The cost of this reduction in variance is that the bias is order one. Only if the regression line is the same for the domain as for the entire population will the bias be zero.

The average mean square error of the three estimators for any subset of small areas can be estimated. If the n_i are small, the estimated variances will provide only limited information for discriminating among estimators. Likewise, there is only one degree of freedom for bias squared for one particular domain. However, a large domain deviation, relative to the standard error, will lead one to reconsider the synthetic estimator.

In their discussion of models, the authors stress the importance of providing estimators of the reliability for small area estimators. They allude to the fact that the principal estimators of mean square error for model based procedures are estimators of an average mean square error. While this is true, it seems worth mentioning that components-of-variance procedures do not assume the mean square errors to be the same in each domain. Also, for the typical survey situation, the estimators of mean square error need not be constant over domains. For example, one of the terms in the mean square error estimator of the components of variance procedure is the estimator of the variance of the direct estimator. The estimated variance of the direct estimator will be a function of the domain sample size and can also be a function of the direct estimated variance of the direct estimator for that domain. See Battese, Harter, and Fuller (1988), Harville (1976), Prasad and Rao (1990), and Ghosh and Rao (1993).

In their discussion of designs, the authors explain that the variance function is often relatively flat in the vicinity of the optimum allocation to strata. A slight reallocation of sample among strata can markedly increase the efficiency of domain estimators for a relatively small decrease in the efficiency of the overall estimates. The same is true with respect to the combination of direct and synthetic estimators. Thus, if one has a relatively good idea of the variance component associated with small areas, either from a previous study on the same population or from a study on a similar population, and if one is under pressure to produce estimates in a brief time span, then it is reasonable to assign fixed weights to form the linear combination. The loss in efficiency is apt to be modest and the programming required for estimation construction considerably reduced. One estimator in this class, and the one adopted by many practitioners, is the synthetic estimator.

The authors briefly raise the question of internal consistency associated with the construction of small area estimates. As they say, if one uses a data dependent procedure, such as variance components, for each dependent variable, then one produces estimates that are not internally consistent. One option is to use multivariate procedures. See, for example, Fuller and Harter (1987) and Fay (1987). Another procedure suggested by Fuller (1990) is to construct components of variance estimators for a limited subset of variables and then use these estimates as control variables in a regression procedure. The regression procedure produces weights for the individual observations. Once the weights are constructed, any number of output tables can be constructed and all estimates are internally consistent.

It is my observation that the gains made in most practical domain estimation problems come primarily from the wise use of auxiliary information. Thus, effort directed towards obtaining quality auxiliary information is effort well spent. If we are able to find a variable x that is highly correlated with the variable y , then there is less variability remaining to be allocated between area to area variance and sampling variance.

ACKNOWLEDGEMENTS

I thank Jay Breidt for comments.

REFERENCES

- BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

- FAY, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*. (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley.
- FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.
- FULLER, W.A., and HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. In *Small Area Statistics*. (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: Wiley.
- HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *The Annals of Statistics*, 4, 384-395.
- GHOSH, M., and RAO, J.N.K. (1993). Small area estimation: An appraisal. Unpublished manuscript. Carleton University, Ottawa, Ontario, Canada.
- PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

COMMENT

GRAHAM KALTON¹

As Singh, Gambino and Mantel (SGM) indicate, there is a growing demand for surveys to provide domain estimates for domains of various sizes and types. This demand is being experienced in many countries throughout the world. In part it may simply reflect a natural growth in the sophistication of survey analysts, who once were content with national estimates and estimates for a few major domains, but who now want to compare and contrast estimates for many different types of domain. In part it results from the needs of policy makers, who require domain information in order to examine how current policies affect different domains, to predict what effects changes in policies might have, and for policy implementation. Information on administrative area domains (*e.g.*, provinces or states, counties, and school districts) is of particular interest for policy purposes (*e.g.*, for identifying low income areas for government support).

In some circumstances the need for domain estimates of adequate precision can be satisfied within the design-based inference framework that is standardly used in the analysis of survey data. This holds for large domains for which the sample sizes are adequate to give the precision required. It can also hold for small domains provided that they are identified in advance, and the sample design is constructed in a way that provides adequate sample sizes. Thus, for example, in the United States, the National Health and Nutrition Examination Survey and the Continuing Survey of Food Intakes by Individuals use differential sampling fractions by age, sex and race/ethnicity and by age/sex and low income status, respectively, in order to provide adequate samples for the domains created by the cross-classifications of these variables. The U.S. Current Population Survey employs differential sampling fractions across the states in order to be able to produce state-level employment estimates. The limitation of this approach is evident when there is a large number of small domains, in which case the sum of the required sample sizes for each domain produces an extremely large overall sample size. This situation occurs often with small administrative districts, such as counties, school districts, and local employment exchanges. In such cases, it may be necessary to discard the standard design-based inference approach in favor of a model-dependent approach that employs a statistical model in the estimation process to borrow strength from data other than that collected in the survey for the given small area. The model-dependent approach may also be required for unplanned small domains, where the need for oversampling had not been foreseen at the design stage.

In response to the demand for small area estimates, a sizeable literature has developed on model-dependent small area estimation methods. Little has, however, been written on the broader issues of small area estimation discussed in the SGM paper, issues that need more attention. Like the authors, I believe that a cautious approach should be adopted to the use of model-dependent small area estimators. I therefore welcome their discussion of methods to make small area estimates within the design-based framework.

From my perspective, the first approach to making small area estimates is to see whether estimates can be produced with adequate precision within the design-based framework. If the domains have been identified in advance, consideration should be given to designing the sample to meet the needs for small area estimates. This may involve ensuring that the small areas do not overlap strata, and ensuring a sufficient sample size for each small area. Another approach suggested by SGM is to minimize the amount of clustering. The smaller the amount of clustering, the less the sample size in each small area is subject to the vagaries of chance. In this regard I see the benefits of less clustering as mainly directed at providing the ability to produce estimates for small areas that were not identified at the design stage. When small areas for which estimates are planned are made into separate strata, the sample size in each small area should be under adequate control even with a clustered sample (provided that the measures of size used in the PPES sampling are reasonable). However, even with planned estimates, there will often be an issue of how to compute variance estimates for a small area from a clustered design, since the number of PSUs sampled in each small area is likely to be small. A variance estimate based on the PSUs within the small area will then be imprecise, with few degrees of freedom, and a generalized variance function approach may be preferred (*e.g.*, assuming that the national design effect applies for each small area). In other words, although the estimate itself may be a design-based estimate, the estimate of its variance may be an indirect one, borrowing strength from other areas. This consideration favors as unclustered a design as possible even for planned small area estimates. The need to model variances is, however, of lesser concern than the need to model the estimates themselves.

An integral part of the design-based framework is a recognition that auxiliary information available for the population may be used at the design stage, at the analysis stage, or at both stages. When information on auxiliary

¹ Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

variables that are closely related to the survey variable is available, substantial gains in precision can accrue. The use of auxiliary information at the analysis stage, through such techniques as post-stratification and ratio, regression and difference estimation, has a special appeal for small area estimation. It should be emphasized that ratio and regression estimators may be motivated by assumptions about the model relating the survey variable (Y) and the auxiliary variables (X), but that the resultant estimators are design-consistent irrespective of the appropriateness of the model. The use of an appropriate model produces the greatest gains in precision, but the estimates are approximately unbiased whatever model is chosen. This may be seen in a simple case where variables X_1, X_2, \dots, X_p are known for every element in the population, and the linear combination $\tilde{Y}_i = B_0 + B_1X_{1i} + \dots + B_pX_{pi}$ is used to estimate Y_i , the value of the Y -variable for population element i . Assume, for simplicity that the B 's are determined from external data, not dependent on the sample. With $Y_i = \tilde{Y}_i + e_i$, the domain total is $Y_a = \sum_{i \in a} \tilde{Y}_i + \sum_{i \in a} e_i = \tilde{Y}_a + E_a$. Since \tilde{Y}_a is known, the estimation problem is one of estimating E_a . From a sample of elements in domain a , E_a may be estimated by $\hat{E}_a = \sum_{j \in s_a} e_j / \pi_j$, where π_j is the selection probability for element j in the sample. The estimator \hat{E}_a is unbiased, independent of the validity of the model employed. The estimation procedure in fact translates the estimation problem from one of estimating Y_a directly to one of estimating E_a and adding on a known constant \tilde{Y}_a . To be effective, the procedure requires the domain variance of the e_i to be smaller than that of the Y_i . There is no requirement that $E_a = 0$. The general logic remains the same in the more usual situation where the B 's are estimated from the sample. In this case, the estimate of Y_a is design-consistent, irrespective of the model adopted (Särndal 1984). Moreover, the B 's may be estimated from the sample data only for the domain of interest, producing what SGM term a direct estimator, or from the total sample, producing a modified direct estimator. A key consideration in the choice between the direct and modified direct estimators in this case is whether the overall B 's also apply for the domain. If not, interaction terms between the X 's and the domain indicators are called for in the total sample model. With a full set of these interaction terms, the modified direct estimator in effect then reduces to the direct estimator.

The need for a model-dependent approach occurs when the design-based estimate lacks sufficient precision even after the auxiliary data available have been used in as effective a manner as possible. Indeed, in some cases the computation of a direct estimate may be impossible because there are no sample cases in the small area. In such situations, it becomes necessary to use a statistical model to borrow strength from other data, often data from other areas. Such models are built upon assumptions (e.g., $E_a = 0$ in the above example), and the quality of the

resultant small area estimates depends on the suitability of the assumptions made. The assumptions are inevitably incorrect to some degree, leading to biases in the small area estimates. Since indirect estimates are biased, the design-based mean square error (MSE) is widely used as the measure of their quality, where $\text{MSE} = V' + B^2$ and V' is the variance and B is the bias of the estimate.

The common way to compare the quality of a direct and an indirect estimate is to compare the variance, V , of the former with the MSE of the latter. However, reading the paper caused me to question whether the MSE is the appropriate measure of quality of an indirect estimator. In a practical setting the variance V of the direct estimate can be estimated whereas the design-based MSE of the indirect estimate cannot. In view of this situation, if $V = \text{MSE}$, then the direct estimator would be clearly preferred. In fact, the direct estimator may tend to be preferred if the direct estimator has adequate precision, irrespective of the likely relative magnitudes of V and MSE. In other cases, if B is the expected bias, then the direct estimator may be preferred to the indirect estimator unless $V > V' + kB^2$, where k is a multiplier greater than 1 that allows for the fact that the unknown bias may be larger than expected.

The same argument can be applied to combined (or composite) estimators that employ a weighted average of a direct and an indirect estimator. Often the principle for choosing the weights is taken to be to minimize the mean square error of the combined estimator, leading to weights for the direct and indirect estimators that are inversely proportional to V and MSE, respectively. However, following the above argument, an alternative procedure would be to minimize the weight of the indirect estimator, subject to the condition that the combined estimator is sufficiently accurate. Alternatively, the weights could be determined on some maximum likely value of the MSE, rather than the expected MSE, to reduce the risk of serious bias in the combined estimator.

I do not follow the rationale for the sample size dependent estimators described by SGM in equation (6.11) and (6.12) in general, but under certain assumptions they may be seen to fit in to the logic given above. With an equal probability sample design and $\delta = 1$, these estimators reduce to the direct estimator when the achieved sample size is greater than, or equal to, the expected sample size. If one assumes that the expected sample size gives adequate precision for the small area, this outcome accords with the above reasoning. If the achieved sample size is smaller than expected, the sample size dependent estimator takes a weighted average of a direct and an indirect estimator. If one assumes that the expected sample size is the minimum sample size to give the required precision, this outcome also accords with the above reasoning. If this indeed is the basis of the sample size dependent estimators, then it would seem useful to generalize them to situations where

the expected sample size is not the sample size that just gives the level of precision required.

As has been noted, auxiliary information plays an important role in the production of accurate small area estimates. Such information may be used for improving the precision of design-based estimates or it may be used in the models employed with the model-dependent approach. Ideally auxiliary information that is highly related to the survey variables involved in the estimates is required. The regular compilation of up-to-date auxiliary data for small areas from administrative and other sources can provide a valuable resource for a small area statistics program.

Although the paper mentions the more general problem of small domains, it focuses predominantly on small areas. This is in line with the general literature and the application of indirect estimation procedures. In part, this may be because the number of socio-economic and other small domains of interest (*e.g.*, age/sex domains) is usually relatively small, compared with the numbers of small areas, so that socio-economic domains can be handled by designing the sample to provide direct estimates of adequate precision for each of them. In part, it may be because the definitions of socio-economic and demographic domains are often chosen in the light of the feasibility of producing design-based estimates of adequate precision for them (*e.g.*, using wider age groupings for some domains); in the case of areal domains, however, the areas are predefined, and no collapsing of areas is acceptable. In part, it may be because there is a lack of auxiliary data to use in the statistical models for such domains. In part, it may also be because the analysis of socio-economic domains is often conducted to make comparisons between the domains. Such comparisons are distorted when the estimate for one

domain borrows strength from other domains (see, for example, Schaible 1992). This issue brings out the general point that indirect estimates should not be uncritically used for all purposes.

In conclusion, I should like to express my support for the general approach of this paper. Where possible, samples should be designed to produce direct small area estimates of adequate precision, and sample designs should be fashioned with this in mind. Auxiliary data should be used, where possible, to improve the precision of direct small area estimates. When indirect estimates are called for, a cautious approach should be used. Models should be developed carefully, estimators that are robust to failures in the model assumptions should be sought, and evaluation studies should be conducted to assess the adequacy of the indirect estimates. Lacking good measures of quality for individual indirect estimates, such estimates need to be clearly distinguished from design-based estimators. Since indirect estimates are not universally valid for all purposes, users need to carefully assess whether the given form of indirect estimate will satisfy their particular needs.

REFERENCES

- SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Design*, (Vol. 1), 95-114, Central Statistical Office of Poland, Warsaw.

RESPONSE FROM THE AUTHORS

We would like to thank Wayne Fuller and Graham Kalton for their stimulating comments, which we find to be quite complementary to the position developed in our paper. In many cases their comments make certain points clearer and strengthen the arguments presented. Encouraged with this kind of endorsement we would like to carry some of the points about survey design further, while responding to the main points made by the discussants.

There is no doubt that survey designers try to optimize the design under operational constraints to meet the stated objectives of a survey. There are usually several objectives to be met by major surveys and it is quite likely that designers have limited influence in the setting of priorities among the various competing objectives. Nevertheless, it is at this stage of priority setting that the case for small area needs should be made strongly, particularly for major continuing surveys.

During the sixties and seventies emphasis in most countries was placed on sub-national (state/provincial) estimates and certain compromises were made to the earlier designs that optimized national estimates. For example, different sampling fractions were used to ensure a minimum sample size for smaller states/provinces. With the demands for data at the sub-state/province level, such as, county, district and municipality, more compromises to the national optimum allocation become necessary, requiring differing sampling fractions among the administrative areas within states/provinces. For example, if the aim is to produce sub-provincial estimates of comparable quality, then provinces will likely receive sample roughly proportional to the number of subprovincial regions they contain. Such an allocation may not be the same as one using the relative population sizes of the provinces. As we discussed in section 5.4, the allocation approach should put more emphasis on a bottom-up strategy. Losses at higher levels and gains at lower levels would differ from survey to survey but it is likely that in many cases a minor loss in CV at the national level will lead to appreciable gains at small area levels.

Kalton stresses the importance of reduced clustering for variance estimation; it is advantageous to increase the degrees of freedom by having a large number of smaller clusters rather than a small number of larger clusters. We would like to emphasize that clustering has another drawback for estimation, and especially small area estimation, namely, a highly clustered design will lead to high design effects, even for planned small domains. The usual reason for resorting to clustered designs is to reduce survey costs. In light of the changes that continue to occur in the data collection process, such as decreased reliance on at-home interviews and increased use of computer assisted interviewing, a periodic review of the cost-variance models that underlie clustering decisions is necessary.

One other issue not addressed in our paper is the impact of sample rotation in continuous surveys. For a given time point, there may be insufficient sample in some small domains to produce reliable estimates. But, as units rotate out of the sample and are replaced, the accumulated or effective sample in the domains increases and may allow the computation of reliable, albeit time-biased, domain estimates. By judicious choice of rotation schemes, survey designers can maximize the cumulative sample size over some time period. For example, for quarterly estimates in a monthly survey, the optimal rotation pattern is $[1(2)]^k$, *i.e.*, repeat the sequence "one month in sample, two months out" k times. This thinking is in the same spirit as Leslie Kish's ideas on cumulation of samples over time.

Kalton clarifies and elaborates the cautious approach to the use of indirect estimators by suggesting a weighted mean squared error, which attaches a weight greater than 1 to the bias term, to allow for the fact that the bias of the indirect estimator may be larger than expected. There are two distinct reasons why the bias may be larger than what is expected from the model for small area effects: random variation within the model, and model breakdown. It is worth recalling here the suggestion of Fay and Herriot (1979) to constrain a combined estimate to be within one standard error of a design estimate; this approach makes allowance for the possibility of large bias in the model estimator for whatever reason. Kalton also reiterates our position that if a direct estimator is of acceptable quality, then in practice, one may decide to use this direct estimator even though its estimated mean squared error exceeds that of model-based competitors. Because there is always the possibility of model failure lurking in the background, this "better safe than sorry" approach is desirable, at least until some experience with particular indirect estimators in specific situations has been gained. This does not contradict the view that there arise situations in which it is necessary to throw caution to the wind.

In his remarks on the sample size dependent estimator, Kalton's comments imply that there is a risk in the strategy which gives the synthetic component zero weight if the observed sample size in the small domain exceeds the expected sample size there since the latter may be too small to yield adequate direct estimates. One option is to use a value n_{\min} which is the size that produces direct estimates that are just barely acceptable. Note, however, that n_{\min} as defined here is characteristic-dependent.

In his comments, Fuller briefly describes an approach to small area estimation that takes advantage of a variance components model and yet has fixed weights for internal consistency among estimators for different characteristics. Besides internal consistency of small area estimates for different characteristics, a second type of consistency that

is sometimes required is that estimates of totals for the set of small areas within a larger area should add up to the published direct estimate for the larger area. One way to achieve this is to benchmark the small area estimates to the direct estimate for the larger area using, for example, a simple ratio adjustment; however, if the ratio adjustment factors depend on the characteristic then this would destroy the first type of consistency. Both types of consistency could be achieved simultaneously if the direct estimators for the larger area are generalized regression estimators, $\hat{Y}_e + (X - \hat{X}_e)\hat{\beta}$, and the modified direct (Section 6.1 in the paper) estimators $\hat{Y}_{sreg,a} = \hat{Y}_{e,a} + (X_a - \hat{X}_{e,a})\hat{\beta}$ are used for small areas.

As Fuller notes, the average squared bias of an estimator for any subset of small areas can be estimated. Here we would like to stress again that the average bias over a set of small areas is not directly relevant for any particular small area. It is for this reason that we prefer to use, whenever possible, estimators that are approximately design unbiased. When use of a model estimator is unavoidable, serious attempts should be made to find appropriate covariates for which reliable auxiliary information is available in order to minimize the residual bias of the model estimator.

Perhaps due to the obvious timeliness problems associated with census data, neither of the discussants commented on censuses as a source of data for smaller domains. In this context it is worth mentioning that some form of ongoing major post-censal survey replacing or supplementing the

decennial census long-form may be considered. Such a strategy, called rolling samples, is described by Kish (1990); a similar approach, called continuous measurement, is described by Alexander (1994). This approach provides a number of options which are worth investigating as potentially cost effective means of producing timely statistics for smaller domains.

Lastly, we would like to stress that the emphasis we put on keeping domain estimation in mind at the design stage, particularly for medium size domains, in no way undermines the important role of models in estimating for very small domains.

We hope that the general direction of the strategy proposed in the paper, supplemented by the fine points brought out by the discussants, particularly the support and cautions summarized by Kalton in his concluding paragraph, will be helpful to survey designers and researchers in finding solutions appropriate to the particular problems they are dealing with.

ADDITIONAL REFERENCES

- ALEXANDER, C.H. (1994). A prototype continuous measurement system for the U.S. Census of Population and Housing. Document for presentation at the annual meeting of the Population Association of America, Miami, Florida, May 5, 1994.
- KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 63-71.

Small Domain Estimation for Unequal Probability Survey Designs

D. HOLT and D.J. HOLMES¹

ABSTRACT

The problem of estimating domain totals and means from sample survey data is common. When the domain is large, the observed sample is generally large enough that direct, design-based estimators are sufficiently accurate. But when the domain is small, the observed sample size is small and direct estimators are inadequate. Small area estimation is a particular case in point and alternative methods such as synthetic estimation or model-based estimators have been developed. The two usual facets of such methods are that information is 'borrowed' from other small domains (or areas) so as to obtain more precise estimators of certain parameters and these are then combined with auxiliary information, such as population means or totals, from each small area in turn to obtain a more precise estimate of the domain (or area) mean or total. This paper describes a case involving unequal probability sampling in which no auxiliary population means or totals are available and borrowing strength from other domains is not allowed and yet simple model-based estimators are developed which appear to offer substantial efficiency gains. The approach is motivated by an application to market research but the methods are more widely applicable.

KEY WORDS: Synthetic estimation; Design-based estimation; Small area estimation; Model-based estimation; Market shares.

1. INTRODUCTION

This paper is concerned with the common problem of estimating domain totals and means from a disproportionately allocated sample survey. Some domains may be large, in which case the achieved sample size may be large too and design-based (or direct) estimators will be satisfactory. Some domains may be small, in which case the achieved sample size may be small too and design-based (or direct) estimators will be too imprecise for practical use. The methods proposed will be motivated through the example of estimating sales, market shares and market penetrations for products in a market research survey. The domains are particular auto manufacturers or models. However, the general approach is applicable to other disproportionately allocated surveys of businesses or institutions.

The problem is analogous to that of using synthetic estimation for small area estimation (Gonzales 1973; Gonzales and Hoza 1978; Platek *et al.* 1987). Synthetic estimation usually depends on two factors: (i) the use of auxiliary variables in conjunction with population means or totals for each small area (or domain) to improve estimates through poststratification or regression estimation, and (ii) the improvement of estimates by pooling data across the small areas (or domains). In our situation no auxiliary population means or totals are available and, since the essential objective is to compare domains (*i.e.*, manufacturers and particular products), the idea of borrowing strength between these is inadmissible. A class

of synthetic estimators is proposed which uses neither of these two approaches and yet is preferred to the direct survey estimators. The proposed estimators have a simple structure, an interesting interpretation and can be justified under a set of model assumptions which are testable under the general assumption of non-informative survey design.

2. THE MARKET RESEARCH EXAMPLE

Market researchers often estimate the total volume of sales and market shares for each manufacturer of a particular product. We consider the case of autos purchased for company fleet use in a single year. Estimates of totals and market shares are required for each auto manufacturer and for specific models which are widely purchased for fleet use.

The terms 'fleet' and 'company' are each interpreted widely. A fleet car is taken to mean any auto purchased on a commercial as opposed to a private basis, and used in conjunction with a business in the broadest sense. This includes autos purchased for sales representatives which may be purchased in large numbers. It also includes single purchases of luxury cars for company directors and other senior staff of large companies, as well as purchases by small 'companies' such as groups of doctors, or self-employed people such as shop owners. Thus the population of purchasing companies – termed consumers – includes a large number of small companies that purchase only one or two autos every few years.

¹ D. Holt and D.J. Holmes, Department of Social Statistics, University of Southampton, Highfield, Southampton, UK, SO95NH.

In the reference period of one year we define Y_{ki} to be the number of autos of product type k purchased by consumer i . The product type k (the domain) may refer to a specific model of a particular manufacturer, or to all models produced by a manufacturer. Thus, $Y_k = \sum_i Y_{ki}$ is the total number of autos of type k purchased by all consumers. Let Z_i be the total number of autos of any kind purchased by consumer i , and $Z = \sum_i Z_i$ be the total number of auto sales. The market share for product type k is defined as $R_k = Y_k/Z$.

We further define

$$Y'_{ki} = 1 \quad \text{if } Y_{ki} > 0 \\ = 0 \quad \text{if } Y_{ki} = 0$$

and

$$Z'_i = 1 \quad \text{if } Z_i > 0 \\ = 0 \quad \text{if } Z_i = 0.$$

Thus, Y'_{ki} and Z'_i are indicator variables for consumers who purchase product type k and at least one auto of any kind, respectively, in the reference period. The number of consumers that purchase product k is thus given by $Y'_k = \sum_i Y'_{ki}$ and the total number of consumers purchasing at least one auto of any kind is given by $Z' = \sum_i Z'_i$. The market penetration for product k , in terms of the proportion of consumers buying a car of any type in the reference period who buy type k , is given by $R'_k = Y'_k/Z'$.

The four parameters Y_k , R_k , Y'_k and R'_k are all legitimate targets of inference in market research and are defined as finite population parameters; namely, domain totals or ratios of domain totals.

3. THE SURVEY DESIGN AND DIRECT ESTIMATORS

The survey design was based upon two mutually exclusive frames and may be regarded as a simple stratified design with ten strata. The first frame was a register (Dun and Bradstreet) of 35,000 companies, stratified into eight strata on the basis of the number of employees and whether the company was classified as 'manufacturing' or 'distributing'. The second frame was a large register of 1.4 million British Telecom business subscribers, stratified into 'private' and 'commercial' numbers. Note that both private and commercial numbers were business subscribers but commercial numbers were allocated if separate commercial premises were occupied.

Using previous survey data the sample was optimally allocated using Neyman allocation to minimize the variance of the estimator of the total number of autos purchased (Z). Data on auto purchases were collected immediately after the end of the reference year. The strata

sizes $\{N_h\}$ and sample allocations $\{n_h\}$ for strata $h = 1, \dots, 10$ are given in Table 1.

Table 1
Sampling Frame: Sample Size and Weight by Stratum

Stratum (h)	Stratum Size N_h	Sample Size n_h	Weight $\pi_h^{-1} = N_h/n_h$
British Telecom:			
Private	389,445	1,150	338.65
Commercial	1,007,399	7,406	136.02
Dun and Bradstreet:			
Manufacturing			
50-99 employees	6,646	235	28.28
100-499	6,826	1,113	6.13
500-999	992	520	1.91
1,000+	1,110	849	1.31
Distributing			
50-99 employees	8,703	472	18.44
100-499	7,625	1,437	5.31
500-999	1,133	484	2.34
1,000+	1,523	1,117	1.36
Overall	1,431,402	14,783	96.83

The sample is a simple, disproportionately allocated stratified design and the direct estimators and their variances are well known. The stratification results in large differences in sampling weights (1.31 to 338.65) and is useful but far from ideal. Many consumers do not purchase any autos at all in the reference year so that each stratum contains a mixture of zero and non-zero responses. For any particular product k the proportion of zero responses in each stratum is obviously larger.

Table 2 contains the direct survey estimates, estimated standard errors (see Holt and Holmes (1993) for derivation), and coefficients of variation for a selection of products from different auto manufacturers. Products A and B represent all models for two major auto manufacturers. Product C is a single model with a substantial share of the fleet market from manufacturer A. The remaining products have small market shares. Products F and G cater for the executive part of the fleet market. The list is incomplete so that the market shares do not sum to one. Also note that the product categories are not mutually exclusive. In general the survey was judged to perform satisfactorily but it was observed over a period of years that estimates for manufacturers or models with small market shares were unstable. This is best seen in terms of the coefficient of variation which is greater than 0.1 for products with small market shares and can be greater than 0.15 or 0.2 in some cases. This instability also affects the estimates of variance as well as the estimates of total sales or market shares of the products.

Table 2

Direct Survey Estimates, Standard Errors and Coefficients of Variation for Selected Products

Product (<i>k</i>)	Estimating Consumers		Estimating Autos	
	Total \hat{Y}'_k	Penetration \hat{R}'_k	Total \hat{Y}_k	Share \hat{R}_k
A	59,890 (2,651) (.044)	.3843 (.0144) (.037)	270,051 (35,704) (.132)	.3781 (.0315) (.083)
B	34,282 (1,960) (.057)	.2200 (.0117) (.053)	153,518 (8,653) (.056)	.2149 (.0131) (.061)
C	23,363 (1,602) (.069)	.1499 (.0098) (.065)	81,381 (17,559) (.216)	.1139 (.0194) (.170)
D	13,857 (1,311) (.095)	.0889 (.0081) (.091)	25,312 (2,906) (.115)	.0354 (.0039) (.110)
E	9,025 (1,146) (.127)	.0579 (.0072) (.124)	24,370 (7,336) (.301)	.0341 (.0101) (.296)
F	5,125 (676) (.132)	.0329 (.0043) (.131)	13,724 (2,369) (.173)	.0192 (.0030) (.156)
G	7,518 (1,015) (.135)	.0482 (.0064) (.133)	11,031 (1,456) (.132)	.0154 (.0022) (.143)

Row 1: estimate

Row 2: s.e.

Row 3: c.v.

4. A MODEL-BASED APPROACH

Given the sample design there is no prospect of improving the efficiency of the direct survey estimators within the conventional sample survey framework. The usual approaches are through the use of auxiliary information for poststratification, ratio or regression estimation but all of these require knowledge of population means or totals. No such information is available. We turn instead to a model-based approach to provide alternative estimators for the whole range of products.

4.1 Estimating Y'_k : the Number of Consumers Purchasing Product Type k

We consider, initially, the number of consumers who buy product type k . We extend the notation from Y'_{ki} to Y'_{khi} in the obvious way to define the indicator random variable of purchase for product k for consumer i in stratum h . We treat each consumer's decision as the outcome of a Bernoulli trial. Let $P_{k|h}$ be the probability that a consumer in stratum h buys an auto of type k [$P_{k|h} = \text{Prob}(Y'_{khi} = 1)$]. We define the model-based equivalent of Y'_k , the total number of consumers of product k , as

$$\Theta'_k = \sum_h N_h P_{k|h}. \quad (1)$$

Assuming that each consumer's decision is independent the likelihood may be written as the usual product of binomial terms. The maximum likelihood estimators are given by $\hat{P}_{k|h} = n_{kh}/n_h$, and the maximum likelihood estimator of Θ'_k is the familiar stratified sampling estimator

$$\hat{\Theta}'_k(1) = \sum_h \frac{N_h}{n_h} n_{kh} = \sum_h N_h \bar{y}'_{kh}, \quad (2)$$

where n_{kh} is the sample count of consumers in stratum h that buy product k , n_h is the stratum sample size and $\bar{y}'_{kh} = n_{kh}/n_h$ is the sample mean for consumers in stratum h (i.e., the sample proportion of consumers in stratum h who buy product k). This estimator is generally unsatisfactory when the sample size for product k is too small.

Suppose we introduce an additional conditioning factor such that every consumer may be categorized into one of its categories f , $f = 1, \dots, F$, and further extend the definition of the indicator random variable to Y'_{khfi} . These categories f will cut across the strata h and the idea is to define f so that, within any particular category, whether a consumer buys product type k or not is independent of the stratum membership h . In the case of fleet purchases we define a categorization based on the total number of autos owned and operated by each consumer (i.e., the fleet size). A more detailed discussion of the choice of f is given in Section 5.

If N_{hff} , the population counts of consumers in stratum h and fleet size category f , are known then (1) may be extended in the obvious way and the target parameter can now be expressed as

$$\Theta'_k = \sum_h \sum_f N_{hff} P_{k|hff}. \quad (3)$$

Equation (3) is the case of poststratification if $\{N_{hff}\}$ are known, and in this case the additional information will lead to a gain in efficiency (Holt and Smith 1979). When $\{N_{hff}\}$ are unknown we may rewrite the model in terms of two sets of probabilities:

$$Q_{f|h} = \text{Prob} \{ \text{consumer has fleet size } f \mid \text{stratum } h \},$$

$$P_{k|hff} = \text{Prob} \{ \text{consumer buys product type } k \mid \text{stratum } h \text{ and fleet size } f \}.$$

The target parameter may now be expressed as

$$\Theta'_k = \sum_h \sum_f N_h Q_{f|h} P_{k|hff}. \quad (4)$$

To obtain an alternative model-based estimator we make further assumptions about the model parameters. Suppose now that

$$P_{k|h} = P_{k|f} \text{ for all } h. \quad (5)$$

This implies that conditional on the categorization f (the size of the fleet operated by a consumer), the probability of buying product type k is *independent* of the original stratum membership h . Algebraically, the assumption is analogous to that used in synthetic estimation for small area estimation but in that case information is pooled across areas. That form of the assumption is inadmissible in our case. We choose instead pooling across strata within the domain of study. The idea is to choose a conditioning variable which accounts for the marginal association between choice of product and stratum membership.

Using assumption (5) and with the obvious extension of the notation ($n_{kf} = \sum_h n_{khf}$, etc.) it may be shown that

$$\hat{Q}_{f|h} = \frac{n_{hf}}{n_h}, \quad \hat{P}_{k|f} = \frac{n_{kf}}{n_f}$$

and the maximum likelihood estimator of Θ'_k becomes

$$\begin{aligned} \hat{\Theta}'_k(2) &= \sum_h \sum_f N_h \frac{n_{hf}}{n_h} \frac{n_{kf}}{n_f} = \sum_f \tilde{N}_f \frac{n_{kf}}{n_f} \\ &= \sum_h \tilde{N}_f \bar{y}'_{kf}, \end{aligned} \quad (6)$$

where $\tilde{N}_f = \sum_h N_h n_{hf}/n_h$, and $\bar{y}'_{kf} = n_{kf}/n_f$ is the unweighted sample mean for consumers in category f (i.e. the sample proportion of consumers in category f who buy product k).

Thus (6) has the form of a stratified estimator based on the categorization f but with the population sizes in each stratum $\{N_f\}$ unknown. Note that an estimator of this form, but with known $\{N_f\}$, would arise naturally if a stratified sample based on f had been selected. In fact this is **not** so: the sample members of category f are **not** selected with equal probability. However, the parameter assumptions lead to treating the sample in each category f as if it was an equal probability sample since under assumption (5) the sample weights are uninformative and simply lead to efficiency loss when estimating $P_{k|f}$. Hence, although the sampling fractions n_h/N_h are used to estimate $\{N_f\}$ they are not used explicitly in $\hat{P}_{k|f} = n_{kf}/n_f = \bar{y}'_{kf}$. Note that the estimator pools information across strata h , within domain k but **not** between domains (i.e. products).

Note that if n_h/N_h is constant, equation (6) reduces to the usual expansion estimator given by (2), and assumption (5) has not yielded a new estimator. If the sample is disproportionately allocated the assumption leads to the

use of the sampling weights for \tilde{N}_f (where they are needed) but not for estimating $P_{k|f}$ (where they are uninformative given f and assumption (5)).

Equation (5) is a strong set of assumptions, requiring $P_{k|h}$ to be exactly equal to a common value $P_{k|f}$ for all h . In practice, random assumptions such as $P_{k|h} = P_{k|f} + \epsilon_{k|h}$ may be introduced, where $E[\epsilon_{k|h}] = 0$ and $V[\epsilon_{k|h}] = \sigma_{\epsilon}^2$. These assumptions will lead to hierarchical Bayes or empirical Bayes analysis as described in Ghosh and Rao (1994) or Fay and Herriot (1979). These methods are not developed here since the simple form of the model-based estimator would be lost, together with the insight that this provides. In a similar vein the approach of Särndal and Hidiriglou (1989) or Drew, Singh and Choudhry (1982) may be applied to yield sample size dependent estimators without violating the requirement that no information is pooled across domains (products).

We can compare the estimators in (2) and (6) when assumption (5) holds since it may be shown that

$$\begin{aligned} V_{\xi}(\hat{\Theta}'_k(1)) &= \sum_h \frac{N_h^2}{n_h} P_{k|h} (1 - P_{k|h}) \\ &= \sum_h \sum_f \frac{N_h^2}{n_h} Q_{f|h} P_{k|f} \\ &\quad - \sum_h \sum_f \sum_{f'} \frac{N_h^2}{n_h} Q_{f|h} Q_{f'|h} P_{k|f} P_{k|f'}, \end{aligned} \quad (7)$$

where the notation $V_{\xi}(\cdot)$ is used to emphasize that the variance is evaluated with respect to the model-based distribution.

It may also be shown that under assumption (5)

$$\begin{aligned} V_{\xi}(\hat{\Theta}'_k(2)) &= \sum_h \sum_f \frac{N_h^2}{n_h} P_{k|f}^2 Q_{f|h} (1 - Q_{f|h}) \\ &\quad - \sum_h \sum_f \sum_{\substack{f' \\ f \neq f'}} \frac{N_h^2}{n_h} P_{k|f} P_{k|f'} Q_{f|h} Q_{f'|h} \\ &\quad + \sum_h \sum_f \frac{N_h^2}{n_h} \frac{P_{k|f} (1 - P_{k|f}) Q_{f|h}}{\sum_h n_h Q_{f|h}} \\ &\quad \left\{ (1 - Q_{f|h}) + n_h Q_{f|h} \right. \\ &\quad \left. + \frac{[1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q_{f|h}^2]}{\sum_h n_h Q_{f|h}} \right\} \end{aligned} \quad (8)$$

and that $V_{\xi}(\hat{\Theta}'_k(1)) - V_{\xi}(\hat{\Theta}'_k(2)) \geq 0$.

Thus under the additional model assumptions $\hat{\Theta}'_k(2)$ has smaller variance as would be expected. These expressions are model-based variances and no finite population corrections arise. A predictive approach to the unobserved elements in each poststratum would give rise to finite population correction factors.

The maximum likelihood estimator of the market penetration for product type k , R'_k , under assumption (5) is simply given by

$$\hat{\Omega}'_k(2) = \frac{\sum_f \hat{N}_f \frac{n_{kf}}{n_f}}{\sum_f \hat{N}_f \frac{n_{af}}{n_f}} = \frac{\sum_f \hat{N}_f \bar{y}'_{kf}}{\sum_f \hat{N}_f \bar{z}'_f}, \quad (9)$$

where n_{af} is the sample count of consumers in fleet category f that buy an auto of any kind, and $\bar{z}'_f = n_{af}/n_f$ is the sample proportion of consumers in category f who buy an auto of any kind.

4.2 Efficiency of the Model-Based Estimator of Y'_k

To investigate the gain in efficiency of $\hat{\Theta}'_k(2)$ over $\hat{\Theta}'_k(1)$ we consider the efficiency of the model-based estimator, defined by

$$e[\hat{\Theta}'_k(2)] = \frac{V_\xi(\hat{\Theta}'_k(1)) - V_\xi(\hat{\Theta}'_k(2))}{V_\xi(\hat{\Theta}'_k(1))}, \quad (10)$$

for various population structures in which assumption (5) holds.

We consider a population with strata $\{h\}$, stratum sizes $\{N_h\}$ and sample allocations $\{n_h\}$ as given in Table 1, and a conditioning factor with ten categories f ($f = 1, \dots, 10$) of increasing fleet size. We compute the efficiency factor $e[\hat{\Theta}'_k(2)]$ for various combinations of parameter values of $\{Q_{f|h}\}$ and $\{P_{k|f}\}$.

We consider five different structures for $\{Q_{f|h}\}$:

$$(a) Q_{f|h} = \begin{cases} 1 & f = h \\ 0 & f \neq h \end{cases} \quad \text{for } h = 1, \dots, 10.$$

$$(b) Q_{f|h} = \begin{cases} 0.95 & f = h & \text{for } h = 1, \dots, 10 \\ 0.025 & f = h - 1 & \text{for } h = 2, \dots, 10 \\ 0.025 & f = h + 1 & \text{for } h = 1, \dots, 9 \\ 0.05 & h = 1, f = 2 \text{ and } h = 10, f = 9 \\ 0 & \text{otherwise} \end{cases}$$

= Band Matrix (0.025, 0.95, 0.025).

$$(c) Q_{f|h} = \text{Band Matrix } (0.05, 0.90, 0.05).$$

$$(d) Q_{f|h} = \text{Band Matrix } (0.05, 0.10, 0.70, 0.10, 0.05).$$

$$(e) Q_{f|h} = 0.1 \quad \text{for } h = 1, \dots, 10 \\ \text{and } f = 1, \dots, 10.$$

We consider four different structures for $\{P_{k|f}\}$:

$$(i) P_{k|f} = \begin{cases} 0.1 & f = 1, 2 \\ 0 & \text{otherwise.} \end{cases}$$

$$(ii) P_{k|f} = 0.1 - 0.01(f - 1) \quad \text{for } f = 1, \dots, 10.$$

$$(iii) P_{k|f} = 0.1f \quad \text{for } f = 1, \dots, 10.$$

$$(iv) P_{k|f} = 0.5 \quad \text{for } f = 1, \dots, 10.$$

Structure (a) is one where the categorization f coincides with the stratification. In structures (b), (c) and (d), in any particular stratum h the majority of consumers fall into one fleet category ($f = h$) with a few consumers in neighbouring categories (e.g., for (b) and (c) $f = h - 1, h + 1$). Finally, structure (e) implies that, in any stratum h , consumers will be equally likely to fall into any one of the fleet categories $f = 1, \dots, 10$.

Structure (i) for $P_{k|f}$ implies a type of auto that is purchased with a small probability by consumers with small fleet sizes (i.e. that fall in categories $f = 1$ or 2), but not purchased by consumers with large (r) fleet sizes. Structure (ii) suggests a type of auto purchased with small probability which decreases as fleet size increases, whilst structure (iii) implies the reverse. In structure (iv) a popular model is bought with probability 0.5 regardless of the consumer's fleet size.

Table 3 gives the efficiency factor defined in (10) for each combination of structures for $Q_{f|h}$ and $P_{k|f}$ under the disproportionate allocation given in Table 1. Column (a) of the table is the special case where the stratification and the categorization f coincide, and the two estimators $\hat{\Theta}'_k(1)$ and $\hat{\Theta}'_k(2)$ are the same. The table shows that large gains in efficiency (e.g., 70%) can be attained for certain parameter combinations: the weaker the association

Table 3
Efficiency Factors, $e[\hat{\Theta}'_k(2)]$, for Various Combinations
of $Q_{f|h}$ and $P_{k|f}$

		Structure for $Q_{f h}$				
		(a)	(b)	(c)	(d)	(e)
Structure for $P_{k f}$	(i)	0	0.108	0.196	0.355	0.648
	(ii)	0	0.116	0.206	0.391	0.695
	(iii)	0	0.103	0.181	0.387	0.695
	(iv)	0	0.115	0.203	0.391	0.706

between f and h the greater the efficiency gain. Even for structures (c) and (d) where the association between f and h is strong, substantial efficiency gains can be achieved. The structure $Q_{f|h}$ is much more important than $P_{k|f}$ in determining efficiency gain.

In the special case (e) where $Q_{f|h}$ is a constant for all f and h it can be shown that the efficiency factor can be expressed as

$$e[\hat{\Theta}'_k(2)] = \left(1 - \frac{\delta^2}{\bar{P}_{k|f} (1 - \bar{P}_{k|f})}\right) \frac{\sum_h \tau_h N_h^2 / n_h}{\sum_h N_h^2 / n_h}, \quad (11)$$

where

$$\bar{P}_{k|f} = \frac{1}{F} \sum_{f=1}^F P_{k|f} \quad \text{and} \quad \delta^2 = \frac{1}{F} \sum_{f=1}^F (P_{k|f} - \bar{P}_{k|f})^2$$

are the mean and variance of $\{P_{k|f}\}$ over the categories f , and $\tau_h = 1 - n_h/n + O(n^{-1})$. The term in parentheses in (11) lies between 0 and 1 and its value depends on how the $\{P_{k|f}\}$ vary over the categories f . In case (iv) $P_{k|f}$ is constant and so this term is unity. The second term of (11) depends solely on the design, and its value for the sample allocation specified in Table 1 is 0.706.

4.3 Estimating Y_k : the Number of Autos Purchased of Product Type k

The previous approach in Section 4.1 may be extended to the number of purchases. We introduce a further conditioning factor which represents the total number of autos purchased, m , regardless of product type, and we extend the notation in the obvious manner to Y_{khfmi} , the random variable representing the number of autos of product type k purchased by consumer i in stratum h , fleet size f , and buying m autos of any kind. The idea is that the number of purchases of product k is likely to vary depending on the total number of autos purchased. Let

$$S_{m|hf} = \text{Prob}\{\text{consumer buys } m \text{ autos of any kind} \mid h, f\}, \\ m = 0, 1, 2, \dots,$$

$$T_{\ell|hfm} = \text{Prob}\{\text{consumer buys } \ell \text{ autos of type } k \mid h, f, m\}, \\ \ell = 0, 1, \dots, m.$$

The model-based target parameter, equivalent to the total purchases of product k , Y_k , is extended from (4) and may now be expressed as

$$\Theta_k = \sum_h \sum_f \sum_m \sum_\ell N_h Q_{f|h} S_{m|hf} T_{\ell|hfm} \ell. \quad (12)$$

We consider two sets of additional assumptions, the first of which is

$$T_{\ell|hfm} = T_{\ell|fjm} \quad \text{for all } h. \quad (13)$$

These assumptions imply that conditional on fleet size category, f , and the total number of new autos purchased, m , the distribution of the number of autos purchased of product type k is independent of stratum h .

The maximum likelihood estimator of Θ_k under assumptions (13) is

$$\hat{\Theta}_k(2) = \sum_f \sum_m \hat{N}_{fjm} \bar{y}_{kfjm}, \quad (14)$$

where $\hat{N}_{fjm} = \sum_h N_h n_{hfjm} / n_h$, and $\bar{y}_{kfjm} = \sum_\ell \ell n_{kfjm} / n_{fjm}$ is the unweighted sample mean of the number of autos of product type k purchased by consumers of fleet size f that purchased a total of m autos of any kind.

The selection probabilities are used here to provide a weighted estimator of N_{fjm} , the total number of consumers of fleet size f that buy m cars of any kind. The form of the estimator is analogous to that in equation (6). Under the model assumption (13) it may be shown that

$$\begin{aligned} V_\xi(\hat{\Theta}_k(2)) &= \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \mu_{fjm}^2 Q_{fjm|h} (1 - Q_{fjm|h}) \\ &\quad - \sum_h \sum_f \sum_m \sum_{f' \neq f} \sum_{m' \neq m} \frac{N_h^2}{n_h} \mu_{fjm} \mu_{f'm'} Q_{fjm|h} Q_{f'm'|h} \\ &\quad + \sum_h \sum_f \sum_m \frac{N_h^2}{n_h} \frac{\sigma_{fjm}^2 Q_{fjm|h}}{\sum_h n_h Q_{fjm|h}} \\ &\quad \left\{ (1 - Q_{fjm|h}) + n_h Q_{fjm|h} \right. \\ &\quad \left. + \frac{[1 + (2n_h - 3)Q_{fjm|h} - 2(n_h - 1)Q_{fjm|h}^2]}{\sum_h n_h Q_{fjm|h}} \right\}, \end{aligned} \quad (15)$$

where $Q_{fjm|h} = Q_{f|h} S_{m|hf}$, $\mu_{fjm} = E_\xi\{Y_{khfmi}\}$, and $\sigma_{fjm}^2 = V_\xi\{Y_{khfmi}\}$.

In practice, \bar{y}_{kfjm} will be based on very few observations if few customers in fleet size category f purchase exactly m cars. For more stability m may be defined as an ordinal variable by grouping the total number of autos purchased into a small number of categories. In this case assumption (13) implies that the distribution of purchases for product type k is the same within fleet size category f and total

purchase category m . Also, ℓ may be treated as a continuous random variable and distributional assumptions made about ℓ leading to ratio or regression estimators.

A second and even stronger set of parameter assumptions is

$$\begin{aligned} T_{\ell|hfm} &= T_{\ell|fm} \quad \text{for all } h, \\ S_{m|h f} &= S_{m|f} \quad \text{for all } h. \end{aligned} \quad (16)$$

These assumptions imply that conditional on fleet size category, f , the joint distribution of the number of autos purchased of type k and the total number of autos purchased of any kind, m , is independent of the stratum h . In this case the maximum likelihood estimator of Θ_k is given by

$$\hat{\Theta}_k(3) = \sum_f \hat{N}_f \bar{y}_{kf}, \quad (17)$$

where $\bar{y}_{kf} = \sum_{\ell} \ell n_{f\ell} / n_f$ is the unweighted sample mean of the number of autos of product type k purchased by consumers in fleet size f regardless of how many autos the consumer bought in total, and $\hat{N}_f = \sum_h N_h n_{hf} / n_h$ is a weighted estimator of the number of consumers of fleet size f overall. It may be shown that under assumptions (16)

$$\begin{aligned} V_{\xi}(\hat{\Theta}_k(3)) &= \sum_h \sum_f \frac{N_h^2}{n_h} \mu_f^2 Q_{f|h} (1 - Q_{f|h}) \\ &\quad - \sum_h \sum_f \sum_{\substack{f' \\ f \neq f'}} \frac{N_h^2}{n_h} \mu_f \mu_{f'} Q_{f|h} Q_{f'|h} \\ &\quad + \sum_h \sum_f \frac{N_h^2}{n_h} \frac{\sigma_f^2 Q_{f|h}}{\sum_h n_h Q_{f|h}} \\ &\quad \left\{ (1 - Q_{f|h}) + n_h Q_{f|h} \right. \\ &\quad \left. + \left[\frac{1 + (2n_h - 3)Q_{f|h} - 2(n_h - 1)Q_{f|h}^2}{\sum_h n_h Q_{f|h}} \right] \right\}. \end{aligned} \quad (18)$$

If assumptions (16) were plausible then \bar{y}_{kf} would be based on larger sample sizes than \bar{y}_{kfm} in (14) and hence $\hat{\Theta}_k(3)$ would be more stable.

The maximum likelihood estimator of the market share for product type k , R_k , under assumption (16), is given by

$$\hat{\Omega}_k(3) = \frac{\sum_f \hat{N}_f \bar{y}_{kf}}{\sum_f \hat{N}_f \bar{z}_f}, \quad (19)$$

where \bar{z}_f , defined analogously to \bar{y}_{kf} , is the unweighted sample mean number of autos of any kind purchased by consumers in fleet category f .

5. EMPIRICAL RESULTS

5.1 Estimating Consumers

In Section 4.2 the efficiency of $\hat{\Theta}_k'(2)$ was investigated for various population structures when assumption (5) held. Readers may find this measure unconvincing since (5) will not hold in practice. We now use the actual survey data to compute $\hat{\Theta}_k'(2)$ for a particular categorization of the conditioning factor that is defined by a combination of the fleet size *and* whether or not the consumer purchased any autos of any kind for fleet use (see Table 4). Empirical evaluations of synthetic estimators have been carried out by Schaible, Brock and Schnack (1977) and Drew, Singh and Choudhry (1982) in different contexts.

For each of the products A-G listed in Table 2 a χ^2 test was used to test the hypothesis that, conditional on the category of the conditioning factor (f), whether or not a consumer purchases that product is independent of stratum (h). Note that for our example the design is stratified random sampling and standard multinomial assumptions apply. For multistage designs, the standard χ^2 analysis would have to be adjusted by using Rao-Scott adjustments for example. In practice it is difficult to find a categorization f such that conditional independence assumptions (5) hold for every product type. However, for the categorization defined in Table 4 it was found that

Table 4
Definition of the Categories, f , of the
Conditioning Factor

Categories f	Definition of f	
	Fleet Size	Fleet Purchases
1	Any	0
2	1-4	> 0
3	5-8	> 0
4	9-15	> 0
5	16-25	> 0
6	26-50	> 0
7	51-100	> 0
8	101-200	> 0
9	201-550	> 0
10	> 550	> 0

most of the variability in the probability of purchasing a particular product type was explained by the category f of the conditioning factor and very little of the residual variation was due to differences in strata.

The model-based estimates for consumers, $\hat{\Theta}_k'(2)$ and $\hat{\Omega}_k'(2)$, obtained from (6) and (9) respectively, are given in Table 5. The model-based variances may give an optimistic view of the precision of the estimators since they depend on the conditional independence assumptions in the model which may be untrue in practice. Alternatively the usual survey estimate of the p -based variance of the model-based estimator may be derived (see Holt and Holmes 1993). This requires no distributional or conditional independence assumptions of any kind and might be considered a more objective measure. These estimates of standard errors are given in Table 5. Since the estimated standard errors are design-based, they include finite population corrections. [We note here that the model-based standard errors for $\hat{\Theta}_k'(2)$ (not shown in Table 5) were consistently around 10% smaller than the p -based standard errors].

Table 5

Model-Based Estimates with p -Based Standard Errors for Selected Products

Product (k)	Estimating Consumers		Estimating Autos	
	Total $\hat{\Theta}_k'(2)$	Penetration $\hat{\Omega}_k'(2)$	Total $\hat{\Theta}_k(3)$	Share $\hat{\Omega}_k(3)$
A	63,433 (2,230)	.4070 (.0105)	263,511 (13,007)	.3722 (.0048)
B	39,673 (1,587)	.2546 (.0086)	177,067 (9,530)	.2501 (.0046)
C	21,930 (1,142)	.1407 (.0066)	65,357 (3,836)	.0923 (.0027)
D	13,422 (868)	.0861 (.0052)	22,146 (1,351)	.0313 (.0016)
E	7,366 (675)	.0473 (.0041)	15,798 (1,223)	.0223 (.0014)
F	5,826 (492)	.0374 (.0031)	14,398 (1,113)	.0203 (.0012)
G	7,686 (633)	.0493 (.0039)	11,207 (813)	.0158 (.0011)

Row 1: estimate

Row 2: p -based s.e.

Comparing these results with the usual survey results given in Table 2 we find that the standard errors for estimating totals are considerably smaller – around 30-40% smaller for all products except A and B (the major manufacturers) where the reduction is about 15-20%. This pattern is expected since the original survey design was optimal for the total sales of autos and therefore relatively

efficient for products with a large market share. We expect the products with smaller market shares to benefit most from the model-based approach.

For estimating market penetration the reduction in standard error is again about 30-40% with slightly smaller reductions for products A and B.

5.2 Estimating Autos

Table 5 also contains model-based estimates for the total number of autos purchased of type k and the corresponding market share, $\hat{\Theta}_k(3)$ and $\hat{\Omega}_k(3)$ as defined by (17) and (19) respectively, for the same categorization f of the conditioning factor as given in Table 4. P -based standard errors for these estimates are also presented in Table 5.

Comparing with the standard survey estimates given in Table 2 large reductions in standard errors for estimating totals are obtained (40-80%) apart from product type B. Similarly, for estimating the market shares the reduction in standard error is again substantial.

6. DISCUSSION

The model-based estimators are derived using conditional independence assumptions to partition the estimation problem into two components. The first, an estimate of N_f (the number of consumers of fleet size f), makes use of the unequal selection probabilities, whereas the second, an estimate of the proportion of consumers of fleet size f buying product type k (or the average number of autos of product type k purchased by consumers of fleet size f) does not. This can result in a substantial efficiency gain.

If the conditional independence assumptions are invalid then in ordinary design-based terms the estimators will have a residual bias but this may be an acceptable risk to achieve stability of the estimators over the whole product range. For the numerical results in previous sections, only the model-based estimates for product B are outside of the 95% confidence interval based on the direct survey estimator. The conditional independence assumptions will depend on the choice of the categories f , and can be tested using chi-square tests for contingency tables.

Whilst the results in Table 5 show that the design-based standard errors for the model-based estimates are generally smaller than for the direct estimates shown in Table 2, it may be argued that the model-based estimators may be biased and hence provide no gain in terms of mean-squared error (MSE). The bias will arise from the inappropriateness of the conditional independence assumptions (e.g., equation (5)). This is not testable, but a comparison of Tables 2 and 5 can give some insight into the size of bias that would be required to cause the MSE to be the same

for both the direct and the model-based estimators. Consider the estimate of total consumers for product E which is strongly affected by the procedure and hence perhaps most susceptible to bias. The variance (and hence MSE) of the direct estimator is $1,146^2 = 1,313,316$ whereas for the model-based estimator the variance is $675^2 = 455,625$. Hence, the model-based estimate of 7,366 would need a bias of 926 in order for the MSEs to be the same.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments.

REFERENCES

- DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 19-47.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. To appear in *Statistical Science*.
- GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- GONZALEZ, M.E., and HOZA, C. (1978). Small area estimation with application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, 73, 7-15.
- HOLT, D., and HOLMES, D.J. (1993). Small domain estimation for unequal probability survey designs. Working Paper Series, No. 2, Department of Social Statistics, University of Southampton, UK.
- HOLT, D., and SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society, Ser. A*, 142, 33-46.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. (1987). *Small Area Statistics*. New York: John Wiley and Sons.
- SÄRNDAL, C.-E., and HIDIRIGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHAIBLE, W.L., BROCK, D.B., and SCHNACK, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the Social Statistics Section, American Statistical Association*, 1017-1021.

Time Series EBLUPs for Small Areas Using Survey Data

A.C. SINGH, H.J. MANTEL and B.W. THOMAS¹

ABSTRACT

In estimation for small areas it is common to borrow strength from other small areas since the direct survey estimates often have large sampling variability. A class of methods called composite estimation addresses the problem by using a linear combination of direct and synthetic estimators. The synthetic component is based on a model which connects small area means cross-sectionally (over areas) and/or over time. A cross-sectional empirical best linear unbiased predictor (EBLUP) is a composite estimator based on a linear regression model with small area effects. In this paper we consider three models to generalize the cross-sectional EBLUP to use data from more than one time point. In the first model, regression parameters are random and serially dependent but the small area effects are assumed to be independent over time. In the second model, regression parameters are nonrandom and may take common values over time but the small area effects are serially dependent. The third model is more general in that regression parameters and small area effects are assumed to be serially dependent. The resulting estimators, as well as some cross-sectional estimators, are evaluated using bi-annual data from Statistics Canada's National Farm Survey and January Farm Survey.

KEY WORDS: Composite estimation; State space models; Kalman filter; Fay-Herriot estimator.

1. INTRODUCTION

There exists a considerable body of research on small area estimation using cross-sectional survey data in conjunction with supplementary data obtained from census and administrative sources. A good collection of papers on this topic can be found in Platek, Rao, Särndal and Singh (1987). Small area estimation techniques in use in U.S. federal statistical programs are reviewed by the Federal Committee on Statistical Methodology (1993). The basic idea underlying all small area methods is to borrow strength from other areas by assuming that different areas are linked via a model containing auxiliary variables from the supplementary data. It would also be important to borrow strength across time because many surveys are repeated over time. Recently time series methods have been employed to develop improved estimators for small areas; see Pfeiffermann and Burck (1990) and Rao and Yu (1992). It is interesting to note that after the initiative of Scott and Smith (1974) on the application of time series methods to survey data, there has only lately been a resurgence of interest in developing suitable estimates of aggregates from complex surveys repeated at regular time intervals; see *e.g.*, Bell and Hillmer (1987), Binder and Dick (1989), Pfeiffermann (1991), and Tiller (1992).

In this paper we consider some natural generalizations of the best linear unbiased predictor (BLUP) for small areas when a time series of direct small area estimates is available. An important example of the BLUP for small areas is the Fay-Herriot (FH) estimator, which entails smoothing of direct estimators by cross-sectional modelling

of small area totals. The resulting estimators are composite estimators (*i.e.*, convex combinations of direct and synthetic estimators) and are called empirical BLUPs, or EBLUPs, whenever estimates of some variance components are substituted in the BLUPs. The work of Fay and Herriot (1979) represents an important milestone in the field of small area estimation because it is probably the first example of a large scale application of small area estimation by government agencies for policy analysis. With the use of structural models, we derive time series EBLUPs which combine both cross-sectional and time series data. The models underlying the time series EBLUPs were chosen on the basis of general heuristic considerations rather than formal model testing procedures. Formal testing of these types of models with survey data is very difficult and not very much is available. Instead, we begin with a regression model that is reasonable for the larger area, and then allow random small area effects to account for any local deviations from the global model. The regression parameters and random small area effects are allowed to evolve over time according to a state space model that was also formulated heuristically. We have not considered here the problem of mean squared error (MSE) estimation for our estimators. MSEs with respect to the motivating models could be defined and estimated for many of the estimators; however, the focus of this paper is on the performance of the estimators in a repeated sampling framework. MSE estimation is an important and difficult problem, and the availability of reliable MSE estimators could be an important consideration in the choice of estimators.

¹ A.C. Singh and H.J. Mantel, Social Survey Methods Division; B.W. Thomas, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The main purpose of this paper is to compare time series EBLUPs with cross-sectional estimators such as post-stratified domain, synthetic, FH and sample size dependent estimators. In the time series modelling of the direct small area estimates we assume that the survey errors are uncorrelated over time. When survey errors are correlated over time and can be modelled reasonably (*e.g.*, ARMA) the approach of Pfeiffermann (1991) can be used to obtain time series EBLUPs via the Kalman filter. Rao and Yu (1992) obtain EBLUPs for a model, in which the Kalman filter cannot be applied, with survey errors having arbitrary correlation structure over time but being uncorrelated across areas. They also develop second order approximations to, and estimation of, the mean squared error under their model. When a model for the correlated survey errors is difficult to specify it may be possible, using a suitably modified Kalman filter, to get good sub-optimal estimators (Singh and Mantel 1991).

In this paper we report on an empirical study of the efficiency of time series EBLUPs. The study uses Monte Carlo simulations from real time series data obtained from Statistics Canada's biannual farm surveys. The main findings of the study are

- (i) There can be reasonable gains in efficiency with time series EBLUPs over cross-sectional estimators.
- (ii) Within the class of time series methods considered in this paper, introduction of serial dependence in the random small area effects is found to be beneficial.
- (iii) Although any smoothed version of the direct small area estimator is expected to be biased, the time series EBLUPs exhibit less bias than cross-sectional smoothing methods.

Section 2 contains a description of various cross-sectional methods for small area estimation. Time series EBLUPs are described in Section 3 and the details and results of the Monte Carlo comparative study are given in Section 4. Finally, Section 5 contains concluding remarks.

2. METHODS BASED ON CROSS-SECTIONAL DATA

In this section we describe some well known small area estimation methods that use survey data from only the current time. Ghosh and Rao (1994) contains a good survey of various small area estimators.

Let Θ denote the vector of small area population totals Θ_k , $k = 1, \dots, K$. In this section, which deals with methods based on cross-sectional data, we ignore the dependence of Θ on time t for simplicity.

2.1 Method 1 (Expansion Estimator for Domains)

This estimator is given by

$$g_{1k} = \sum_{j \in s_k} d_j y_j,$$

where d_j is the survey weight for sample unit j . For stratified simple random sampling, which is used for our simulation study in Section 4, we have

$$g_{1k} = \sum_h (N_h/n_h) \sum_{j \in s_{hk}} y_{hj}, \quad (2.1)$$

where y_{hj} is the j -th observation in the h -th stratum, s_{hk} denotes the set of n_{hk} sample units falling in the k -th small area in the h -th stratum and n_h , N_h denote respectively the sample and population sizes for the h -th stratum. This estimator is often unreliable because n_{hk} , the random sample size in the small area, may be small in expectation and could have high variability. Conditional on the realized sample size n_{hk} , g_{1k} is biased. However, unconditionally, it is unbiased for Θ_k .

2.2 Method 2 (Post-stratified Domain Estimator)

We will also refer to this estimator as the direct small area estimator. If the population size N_{lk} is known for some post-strata indexed by l , then the efficiency of the estimator g_{1k} could be improved by post-stratification. We define

$$g_{2k} = \sum_l N_{lk} \sum_{j \in s_{lk}} d_j y_j / \sum_{j \in s_{lk}} d_j = \sum_l N_{lk} \bar{y}_{lk}.$$

In our simulations our post-strata are the intersections of design strata with small areas which leads to

$$g_{2k} = \sum_h (N_{hk}/n_{hk}) \sum_{j \in s_{hk}} y_{hj} = \sum_h N_{hk} \bar{y}_{hk}. \quad (2.2)$$

This estimator also may not be sufficiently reliable because of the possibility of n_{hk} 's being small in expectation. If $n_{hk} = 0$, the above estimator is not defined. It is conventional to replace \bar{y}_{hk} by 0 when $n_{hk} = 0$. In the empirical study presented in this paper, we replaced \bar{y}_{hk} by the synthetic estimate $(\bar{X}_{hk}/\bar{X}_h)\bar{y}_h$, where X is a suitable covariable, whenever $n_{hk} = 0$.

The estimator g_{2k} in (2.2) is conditionally (given $n_{hk} > 0$) unbiased and approximately unconditionally unbiased. Appendix A.1 gives details of estimation of the conditional mean squared error, v_k , of g_{2k} .

2.3 Method 3 (Synthetic Estimator)

It is possible to define a more efficient estimator by assuming a model which allows for "borrowing strength" from other small areas. This gives rise to synthetic estimators, see *e.g.*, Gonzalez (1973) and Erickson (1974). Suppose different small area totals are connected via the auxiliary variable X_k by a linear model as

$$\Theta_k = \beta_1 + \beta_2 X_k, \quad k = 1, \dots, K, \quad (2.3a)$$

or in matrix notation

$$\underline{\Theta} = F\underline{\beta}, \quad (2.3b)$$

where $F = (F_1, F_2, \dots, F_K)'$, $F_k = (1, X_k)'$. Now consider a model for the direct small area estimators g_{2k} 's as

$$g_2 = F\underline{\beta} + \underline{\epsilon},$$

where $g_2 = (g_{21}, \dots, g_{2K})'$, $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_K)'$, ϵ_k 's are uncorrelated survey errors with mean 0 and variance v_k . Note that the g_{2k} 's are uncorrelated over areas since they are conditionally (given n_{hk}) unbiased and the samples in different small areas are conditionally independent.

Denoting by $\hat{\underline{\beta}}$ the weighted least squares (WLS) estimate of $\underline{\beta}$, we obtain the regression-synthetic estimator of $\underline{\Theta}_k$ under the assumed model as

$$g_3 = F\hat{\underline{\beta}}.$$

The above estimator could be heavily biased unless the model (2.3) is satisfied reasonably well. The above model may not be realistic because no random fluctuation or random small area effect (a_k , say) is allowed.

2.4 Method 4 (Fay-Herriot Estimator or EBLUP)

Using the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor approach (see *e.g.*, Battese, Harter and Fuller 1988, and Pfeffermann and Barnard 1991), the bias of the synthetic estimator can be reduced considerably by using a composite estimator; for an early reference on composite estimation see Schaible (1978). The composite estimator is obtained as a convex combination of g_2 and a modified g_3 . For this purpose, it is assumed that

$$\underline{\Theta} = F\underline{\beta} + \underline{a}, \quad (2.4)$$

where a_k 's are uncorrelated random small area effects with mean 0 and variance w_k known up to a constant. In our empirical study later we take $w_k = w$. Thus we model g_2 as

$$g_2 = F\underline{\beta} + \underline{a} + \underline{\epsilon}. \quad (2.5)$$

Here \underline{a} is also assumed to be uncorrelated with $\underline{\epsilon}$. The BLUP of $\underline{\Theta}$ under the model defined by (2.4) and (2.5) is

$$\begin{aligned} g_4 &= g_3^* + \Lambda(g_2 - g_3^*) \\ &= \Lambda g_2 + (I - \Lambda)g_3^*, \end{aligned} \quad (2.6)$$

where

$$\Lambda = (V^{-1} + W^{-1})^{-1}V^{-1} = WU^{-1}, \quad U \equiv V + W,$$

$$V = \text{diag}(v_1, \dots, v_K), \quad W = \text{diag}(w_1, \dots, w_K),$$

and $g_3^* = F\underline{\beta}^*$, $\underline{\beta}^*$ is the WLS estimate of $\underline{\beta}$ under model (2.5). Here it is assumed that both the covariance matrices V and W are known in computing the BLUP.

The expression (2.6) follows from the general results on linear models with random effects, see *e.g.*, Rao (1973, p. 267) and Harville (1976). The BLUP or BLUE of $F\underline{\beta}$ is g_3^* and the BLUP of \underline{a} is $\Lambda(g_2 - g_3^*)$. It may be of interest to note that the structure of the BLUP does not change regardless of whether or not $\underline{\beta}$ is known. However, its MSE does change as expected due to estimation of $\underline{\beta}$.

When V and W are replaced by estimates, the estimator g_4 is termed EBLUP. Note that the model (2.4) is more realistic than (2.3), and therefore, the performance of g_4 is expected to be quite favourable. The estimator g_4 approaches g_2 when the v_k 's get small, *i.e.*, when the n_{hk} 's become large. However, it remains biased, in general, conditional on $\underline{\Theta}$, with bias tending to 0 as the v_k 's get small.

2.5 Method 5 (Sample Size Dependent Estimator)

An alternative composite estimator is given by the sample size dependent estimator of Drew, Singh and Choudhry (1982). It is defined as

$$g_5 = \Delta g_2 + (I - \Delta)g_3,$$

where $\Delta = \text{diag}(\delta_1, \dots, \delta_K)$,

$$\delta_k = \begin{cases} 1 & \text{if } \sum_{j \in s_k} d_j \geq \lambda N_k, \\ \sum_{j \in s_k} d_j / \lambda N_k & \text{otherwise} \end{cases} \quad (2.7)$$

and the parameter λ is chosen subjectively as a way of controlling the contribution of the synthetic component. The above estimator takes account of the realized sample size n_{hk} 's and if these are deemed to be sufficiently large according to the condition in (2.7), then it does not rely on the synthetic estimator. This property is somewhat similar to that of g_4 ; however, unlike g_4 , the above estimator does not take account of the relative sizes of the within area and between area variation. Rao and Choudhry (1993) have demonstrated empirically how EBLUPs can sometimes outperform sample size dependent estimators, especially when the between area variation is not large relative to the within area variation. Särndal and Hidirolou (1989) also proposed estimators similar to the above sample size dependent estimator.

3. METHODS BASED ON POOLED CROSS-SECTIONAL AND TIME SERIES DATA

Suppose information is available for several time points, $t = 1, \dots, T$, in the form of direct small area estimators g_{2t} , where g_{2t} is the vector of estimates g_{2k} in (2.2) based on data from time t , and also the small area population totals for the auxiliary variable. We will now introduce some estimators which generalize the Fay-Herriot estimator g_{4T} in different ways by taking account of the serial dependence of the direct estimates $\{g_{2t} : t = 1, \dots, T\}$. Recall that for the Fay-Herriot estimator, the model for Θ_T has two components, namely, the structural component $F_T \beta_T$ and the area component a_T . The estimator g_{4T} borrows strength over areas for the current time T and is given by the sum of two components, each being EBLUP (BLUE) for the corresponding random (fixed) effect, *i.e.*,

$$g_{4T} = F_T \beta_T^* + a_T^*. \quad (3.1)$$

Methods based on time series data could, however, borrow strength over time as well. Here we introduce three estimators which are motivated from specific structural models for serial dependence. All three of these estimators are optimal under different special cases of a structural time series model for the direct small area estimates $\{g_{2t} : t = 1, \dots, T\}$ specified by the following state space model. Let α_t denote $(\beta_t', a_t')'$ and H_t denote (F_t, I) . Then we have

$$g_{2t} = \Theta_t + \xi_t, \quad (3.2a)$$

$$\Theta_t = F_t \beta_t + a_t \equiv H_t \alpha_t$$

and

$$\alpha_t = G_t \alpha_{t-1} + \zeta_t, \quad (3.2b)$$

where

$$G_t = \begin{pmatrix} G_t^{(1)} & 0 \\ 0 & G_t^{(2)} \end{pmatrix}, \quad \zeta_t = \begin{pmatrix} \xi_t \\ \eta_t \end{pmatrix}, \quad (3.2c)$$

along with the usual assumptions about random errors, *i.e.*, ξ_t , ζ_t are uncorrelated, ζ_t is uncorrelated with α_s for $s < t$, and that $\xi_t \sim (0, V_t)$, $\zeta_t \sim (0, \Gamma_t)$ where $\Gamma_t = \text{block diag}\{B_t, Q_t\}$. The covariance matrices V_t , B_t , and Q_t are generally diagonal. If $G_t^{(1)} = I$ and $G_t^{(2)} = I$ then β_t and a_t evolve according to a random walk.

This model is in the general class defined by Pfeffermann and Burck (1991) using structural time series models. The main purpose of their study was to show how accounting for cross-sectional correlations between neighbouring small areas (in addition to serial correlations) and inclusion of certain robustness modifications (to protect against

model breakdowns) could improve the performance of time series model based estimators. They also used the maximum likelihood method under normality to estimate model parameters. The focus of this paper, on the other hand, is on the Monte Carlo evaluation of a special class of time series estimators (related to Fay-Herriot) chosen on the basis of heuristic considerations and not on the basis of model fitting. The methods considered could, therefore, be viewed as model assisted methods whose performance will be evaluated in a design based (*i.e.*, repeated sampling) framework by Monte Carlo simulation. Moreover, it will be seen later that, for the types of serial dependence considered, the model parameters can be estimated relatively simply by the method of moments, without making any distributional assumptions such as normality.

To find the optimal estimator (BLUP) of Θ_T in (3.2) based on all the direct estimates up to time T , we first found the BLUP $\tilde{\alpha}_T$ of α_T from which the BLUP of Θ_T is obtained as $H_T \tilde{\alpha}_T$. It is possible, albeit cumbersome, to get $\tilde{\alpha}_T$ directly from the complete data using the theory of linear models with random effects. However, since the α_T s are connected over time according to the transition equation (3.2b), it is more convenient to compute it recursively using the Kalman filter (KF). Traditionally KF is viewed as a Bayesian technique in which at each time t , the posterior distribution of α_t given data up to $t - 1$ is updated to get the posterior distribution of α_t given data up to time t . Although it is instructive to view KF in this manner, it is not necessary under mixed linear models. Suppose $\tilde{\alpha}_{T|s}$ denotes the BLUP of α_T based on data up to time s , $s < T$. It is known (see Duncan and Horn 1972) that, for the special structure of serial dependence considered here, the BLUP $\tilde{\alpha}_T$ of α_T based on data up to time T is the same as the BLUP of α_T based on $\tilde{\alpha}_{T|s}$ and the last $T - s$ observations. In other words, information in the previous data can be condensed into an appropriate BLUP before augmenting more current data points. A good description of the Kalman filter is given in chapter 3 of Harvey (1989).

3.1 Method 6 (Time Series EBLUP-I)

For the first estimator, we let β_t evolve over time (*e.g.*, according to a random walk), but assume that a_t is serially independent. The equations for the state space model for this case are similar to (3.2) except that the serial independence of the a_t s implies $G_t^{(2)} = 0$. This will give rise to a composite estimator

$$g_{6T} = F_T \tilde{\beta}_T + \tilde{a}_T. \quad (3.3)$$

Note that $\tilde{\beta}_T$ in (3.3) would now be based on all the small area estimates up to time T and therefore would be different from β_T^* of (3.1) which is based on only direct estimates at time T . The estimator \tilde{a}_T , as a result, would also be different from the corresponding component a_T^* of (3.1).

In the simulation study described later we take $G_t^{(1)} = I$, $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$, corresponding to a random walk model, and $Q_t = \tau^2 I$. Appendix A.2 illustrates the method of moments estimation of the parameters γ_1^2 , γ_2^2 , and τ^2 . The KF may then be run, with initial values for $\tilde{\alpha}_1$ and its MSE obtained from the FH estimator at $t = 1$, to obtain the EBLUP of $\tilde{\alpha}_T$. Then $H_T \tilde{\alpha}_T$ is the time series EBLUP-I estimator g_{6T} at time T .

As pointed out by a referee, when the number of small areas is quite large, or when the variation in β_t over t is relatively large, there is little difference between g_{6T} and g_{4T} . Indeed, there is little difference between the performances of these two estimators in our simulation study described in Section 4.

3.2 Method 7 (Time Series EBLUP-II)

For the second estimator, we let β_t be fixed (it may or may not be common for different time points) and let the area effects q_t be serially dependent according to, for example, a random walk. This time series generalization could be viewed as an analogue of the model proposed by Rao and Yu (1992). The resulting composite estimator will have the same form as (3.1), *i.e.*,

$$g_{7T} = F_T \tilde{\beta}_T + \tilde{q}_T, \quad (3.4)$$

but the component estimates $\tilde{\beta}_T$ and \tilde{q}_T would be different. We have two cases.

3.2.1 Case 1: Suppose the β_t s are fixed and time-invariant but the q_t s are serially dependent. Then, in (3.2), $G_t^{(1)} = I$ and $B_t = 0$. If Q_t is taken as $\tau^2 I$, then the only unknown parameter τ^2 can be estimated by the method of moments; see Appendix A.2. We will denote by g_{7T} the EBLUP obtained in this case when the parameter estimate is substituted.

3.2.2 Case 2: Here we assume that β_t s are fixed but different for different time points. The area effects q_t evolve over time as in Case 1. In (3.2) we have $G_t^{(1)} = 0$ and $B_t = mI$ where m is a large number. The expressions for \tilde{q}_T and its MSE obtained from the KF in this case give the correct formulas as $m \rightarrow \infty$ (see Sallas and Harville 1981). The KF updating equations for \tilde{q}_t in this case take the special form

$$\begin{aligned} \tilde{\beta}_t &= (F_t' A_t^{-1} F_t)^{-1} F_t' A_t^{-1} (g_{2t} - G_t^{(2)} \tilde{q}_{t-1}); \\ \tilde{q}_t &= G_t^{(2)} \tilde{q}_{t-1} + P_{t|t-1} A_t^{-1} (g_{2t} - G_t^{(2)} \tilde{q}_{t-1} - F_t \tilde{\beta}_t); \\ P_t &= P_{t|t-1} - P_{t|t-1} A_t^{-1} (A_t - F_t (F_t' A_t^{-1} F_t)^{-1} F_t') \\ &\quad A_t^{-1} P_{t|t-1}, \end{aligned}$$

where $A_t = P_{t|t-1} + V_t$, P_t is the MSE of \tilde{q}_t about q_t , and $P_{t|t-1} = G_t^{(2)} P_{t-1} \{G_t^{(2)}\}' + Q_t$ is the MSE of $G_t^{(2)} \tilde{q}_{t-1}$ as an estimator of q_t . The time series EBLUP in this case will be denoted by g_{7T}^* .

3.3 Method 8 (Time Series EBLUP-III)

For the third estimator, we let both β_t and q_t evolve over time. This will have more complex serial dependence than either (3.3) or (3.4). Its form will be similar to (3.1) and can be represented as

$$g_{8T} = F_T \tilde{\beta}_T + \tilde{q}_T. \quad (3.5)$$

As before, if $B_t = \text{diag}\{\gamma_1^2, \gamma_2^2\}$ and $Q_t = \tau^2 I$, then the model parameters τ^2 , γ_1^2 , γ_2^2 can be estimated by the method of moments as in Appendix A.2. The resulting EBLUP of Θ_T will be denoted by g_{8T} .

It may be of interest to note that many of the estimators considered so far are optimal under special cases of the model underlying g_{8T} . As has been shown, the time series EBLUPs of methods 6 and 7 result from making restrictions on the matrices G_t and Γ_t . The cross-sectional Fay-Herriot estimators of Section 2.4 result from restricting the data to a single time point. The synthetic estimators of section 2.3 are special cases of the Fay-Herriot estimators with zero variance for the random small area effects, and the direct (post-stratified) estimator is obtained in the limit as the variance of the small area effects goes to infinity.

A further generalization that could be useful is to allow correlations between neighbouring small area effects. This can be accomplished by allowing the matrix Q_t in (3.2) to be non-diagonal; however, it is not clear what would be an appropriate correlation structure in Q_t .

4. MONTE CARLO STUDY

The cross-sectional and time series methods were compared empirically by means of a Monte Carlo simulation from a real time series obtained from Statistics Canada's biannual farm surveys, namely, the National Farm Survey (in June) and the January Farm Survey. Due to the redesign after the census of Agriculture in 1986, the survey data for the six time points starting with the summer of 1988 were employed to create a pseudo-population for simulation purposes. To this, data from the census year 1986 was also added. Thus information at one more time point was available although this resulted in a 3-point gap in the series. The missing data points, however, can be easily handled by time series methods. It may be noted that although the data series is short, it is nevertheless believed to be adequate for illustrative purposes. The parameter of interest was taken as the total number of cattle and calves for each crop district (defined as the small area) at each time point. For simplicity, independent stratified random samples were drawn for each occasion from the pseudo-population, though the farm surveys use rotating panels over time. The dependence of direct small area estimates over time was modelled by assuming that the underlying

small area population totals are connected according to some random process. The auxiliary variable used in the model was the ratio-adjusted census 1986 value of the total cattle and calves for each small area. This showed high correlations with the corresponding variable over time at the farm level. Specific details of the empirical study are described below.

4.1 Design of the Simulation Experiment

First we need to construct a pseudo-population from the survey data over six time points (June 1988, January 1989, . . . , January 1991). The actual design involves two frames (list and area) with a one stage stratified sampling from the list frame and a two stage stratified sampling from the area frame, for details see Julien and Maranda (1990). We decided to use survey data from the list frame only because the list frame corresponds to farms existing at the time of Census 1986 and the chosen auxiliary variable for model building was based on Census 1986 information. Moreover, we chose to use the data from the province of Quebec because its area sample is only a minor component of the total sample and the estimated coefficient variation for the twelve crop-districts (*i.e.*, small areas of interest) of this province showed a wide range for the livestock variables. It was decided to avoid variability due to changes in the underlying population over time by retaining only those farms which responded to all the six occasions. Also, farm units who belonged to a multiholding arrangement in any one of the seven time points (including the census) were excluded because of the problems in finding individual farm's data from the multiholding summary record and changes in their reporting arrangement over time.

The various exclusions described above were motivated from considerations of yielding a sharper comparison between small area estimators. The total count of farm units after exclusions was found to be 1,160 out of a total of over 40,000 farms on the list frame. For the pseudo-population, we replicated the 1,160 farm units proportional to their sampling weight so that the total size N of the pseudo-population was 10,362, which was manageable for micro-computer simulation.

The pseudo-population was stratified into four take-some and one take-all strata using Census 1986 count data on cattle and calves as the stratification variable. Although we did not consider alternative stratifications or sample sizes in our simulation study, there is no reason to think that our conclusions would alter significantly if we were to do so. The sigma-gap rule (Julien and Maranda 1990) was used for defining the take-all stratum. To apply the sigma-gap rule we look at the smallest population value greater than the population median where the distance to the next population value, in order of size, is at least one population standard deviation; all units above this point are placed into the take-all stratum. The algorithm of Sethi

(1963) was used for determining optimal stratification boundaries for take-some strata. Neyman's optimum allocation was used for sample sizes for strata in order to optimize the precision of the provincial estimate of total count. This resulted in, from a total sample size of 207 (2% sampling rate), allocations of 51, 62, 48 and 35 from takesome strata with 5,001, 3,188, 1,850 and 312 farms, respectively, and the size of the take all stratum was 11. The expected number of sample farms in each small area varied from 4.6 in area 9 up to 27.5 in area 6, with an average of 17.3. The expected number of sample farms with some cattle and calves varied from 3.6 in area 9 to 18.8 in area 3, and the average over the small areas was 11.7. A total of 30,000 simulations were performed. For each simulation, samples were drawn independently for each time point using stratified simple random sampling without replacement. The 30,000 simulations were conducted in 15,000 sets of 2 simulations where each set corresponds to a different vector of realized sample sizes in the twelve small areas within each stratum. This was required to compute certain conditional evaluation measures as described in the next subsection, see also Särndal and Hidiroglou (1989).

4.2 Evaluation Measures

Suppose m simulations are performed in which m_1 sets of different vectors of realized sample sizes in domains (h, k) are replicated m_2 times. The following measures can be used for comparing performance of different estimators at time T . Let i vary from 1 to m_1 and j from 1 to m_2 .

(i) Absolute Relative Bias for area k :

$$ARB_k = |m^{-1} \sum_i \sum_j (est_{ijk} - true_k) / true_k|. \quad (4.1)$$

The average of ARB_k over areas k will be denoted by $AARB$. We take the absolute relative bias since our primary interest in this study is in an overall measure like $AARB$; however, in other contexts the actual biases for individual small areas may also be of considerable interest.

The following measure is motivated by a desire to evaluate the conditional performance of estimators, conditional on the vectors of realized sample sizes in domains. It is conventional to measure performance conditional on fixed domain sample sizes; here we consider the standard deviation of the conditional bias, B_{ik} , as a simple summary measure. If this standard deviation is small then the method is robust to variations in the realized sample sizes. Note that the expected value of B_{ik} is just the unconditional bias which is estimated by ARB_k . Let B_k^2 denote the unconditional expected value of B_{ik}^2 . We define the following Monte Carlo measure:

- (ii) Standard Deviation of Conditional Relative Bias for area k :

$$\text{SDCRB}_k = \left\{ m_1^{-1} \sum_i (\hat{B}_{ik}^2 - \hat{C}_{ik}) / \text{true}_k - \text{ARB}_k^2 \right\}^{1/2};$$

$$\hat{B}_{ik} = m_2^{-1} \sum_j \text{est}_{ijk} - \text{true}_k, \quad (4.2)$$

$$\hat{C}_{ik} = m_2^{-1} (m_2 - 1)^{-1} \left(\sum_j \text{est}_{ijk}^2 - \left(\sum_j \text{est}_{ijk} \right)^2 / m_2 \right).$$

The correction term \hat{C}_{ik} adjusts for bias in \hat{B}_{ik}^2 , as an estimate of B_{ik}^2 , due to m_2 being finite. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ is conditionally unbiased for B_{ik}^2 ; it is also unconditionally unbiased for B_k^2 . The Monte Carlo average $m_1^{-1} \sum_i (\hat{B}_{ik}^2 - \hat{C}_{ik})$ converges to B_k^2 with probability 1 as $m_1 \rightarrow \infty$. $\hat{B}_{ik}^2 - \hat{C}_{ik}$ may be negative for some i , due to finite m_2 . For large m_1 the average over i is usually very close to B_k^2 ; whenever the average is less than ARB_k^2 we set SDCRB_k to 0. ASDCRB will denote the average of SDCRB_k over areas k .

- (iii) Mean Absolute Relative Error for area k :

$$\text{MARE}_k = m^{-1} \sum_i \sum_j | \text{est}_{ijk} - \text{true}_k | / \text{true}_k \quad (4.3)$$

and AMARE denotes the average of MARE_k over areas.

- (iv) Mean Squared Error for area k :

$$\text{MSE}_k = m^{-1} \sum_i \sum_j (\text{est}_{ijk} - \text{true}_k)^2 \quad (4.4)$$

and AMSE as before denotes the average over areas.

- (v) Relative Root Mean Squared Error for area k :

$$\text{RRMSE}_k = \{ \text{MSE}_k \}^{1/2} / \text{true}_k. \quad (4.5)$$

Again, ARRMSE denotes the average over areas.

The precision (*i.e.*, the Monte Carlo standard error) of each measure depends on m_1 , m_2 . For all measures except (ii), the optimal choice of m_1 , m_2 under the restriction that $m_2 > 1$ is $m_1 = m/2$, $m_2 = 2$, since this minimizes the Monte Carlo standard error. To see this, let A be the average of an evaluation measure from m_2 samples all with the same sample configuration (set of random sample sizes in domains) which we call C . Then the expected value of A conditional on C is a function of C ,

say $E(C)$, and the conditional variance of A is proportional to m_2^{-1} , say $V(C)/m_2$. The unconditional variance of A is then $V\{E(C)\} + E\{V(C)\}/m_2$, and the overall Monte Carlo variance of an evaluation measure based on m_1 sample configurations replicated m_2 times is $V\{E(C)\}/m_1 + E\{V(C)\}/m_1 m_2$ which is minimized, since $m = m_1 m_2$ is fixed, by taking m_1 as large as possible. For the second measure, the appropriate choice of m_1 , m_2 is less straightforward. In the simulation study, m was chosen as 30,000 and the corresponding values of m_1 , m_2 were set at 15,000 and 2.

4.3 Estimators Used in the Comparative Study

There were nine estimators included in the study, namely, g_1 to g_8 and g_7^* , all calculated for time $T = 10$. We used a simple linear regression model for the synthetic component with the auxiliary variable defined as

$$X_{kt} = (\hat{\Theta}_t / \Theta_1) \Theta_{k1}, \quad (4.6)$$

where Θ_{k1} , Θ_1 respectively denote the population totals for small area k and the province at $t = 1$, *i.e.*, at Census 1986. The estimator $\hat{\Theta}_t$ denotes the post-stratified estimator of Θ_t from the farm survey at time t at the province level. Thus X_{kt} is simply a ratio-adjusted synthetic variable. The variances of error components in the regression model were assumed to be constant over areas. For time series models, it was assumed that the serial dependence was generated by a random walk. The above type of model assumptions have been successfully used in many applications and the main reason for our choice was simplicity. It was hoped, however, that the chosen models might be adequate for our purpose and might illustrate the differential gains with different types of model assisted small area estimators, *i.e.*, both cross-sectional and time series smoothing methods.

Since the Census 1986 data was included in the time series, the direct estimate g_{21} corresponds to Census 1986 and therefore the survey error ϵ_{11} would be identically 0. Moreover, from the definition of X_{kt} , it follows that a reasonable choice of (β_{11}, β_{21}) would be $(0, 1)$ which implies that g_1 must be 0. Thus the covariance matrices B_t and W_t at $t = 1$ are null and, therefore, the distribution of α_t at $t = 1$ would not require estimation. The above modification in the initial distribution of α_t is natural in view of the extra information available from the census. Moreover, since the direct estimates g_{2t} were not available for $t = 2, 3, 4$, equations for estimating model variance components in Appendix A.2 were modified accordingly.

For method 7 (case 1), β_t was assumed to have a common fixed value only for $t \geq 2$ because at $t = 1$, $\beta_t = (0, 1)'$. For the sample size dependent estimator g_5 the parameter λ was taken to be 1.

4.4 Empirical Results

The main findings were listed in Section 1. Here we give some detailed comparisons and some possible explanations. We do not show separate results for g_7^* which performs slightly worse than, though overall similarly to, g_7 . The estimators are summarized in Table 1. Figures 1 to 3 and Tables 2 to 4 present some of the empirical results. We have not shown the Monte Carlo standard errors but they were all found to be quite negligible.

Table 1
Summary of Estimators

g_1 – Expansion	g_6 – Time Series EBLUP-I, β s evolve over time, as independent over time
g_2 – Post-stratified	
g_3 – Synthetic	g_7 – Time Series EBLUP-II, as evolve over time, fixed common β
g_4 – Fay-Herriot	
g_5 – Sample Size Dependent	g_8 – Time Series EBLUP-III, β s and as evolve over time

Table 2 gives the five evaluation measures averaged over small areas, Figure 1 shows plots of the averaged evaluation measures relative to the Fay-Herriot (g_4) value. There is a clear pattern in the behaviour of various measures across different estimators. The direct estimator g_2 does very well with respect to the bias measure (AARB) but does somewhat poorly with respect to the other measures. The cross-sectional smoothing method g_3 (synthetic) does quite poorly with respect to the bias measures. The Fay-Herriot method g_4 performs somewhat better than post-stratified on average with respect to the MSE measure but is much worse in terms of bias. The sample size dependent method g_5 is quite similar to g_2 , slightly worse with respect to the bias measures and slightly better with respect to the other measures. The time series methods g_7 and g_8 perform quite well overall, though they are somewhat worse than g_2 with regard to bias. The performance of the time series estimator g_6 is generally between that of Fay-Herriot and the time series estimators g_7 and g_8 . For all of the estimators (including the synthetic g_3) the standard deviation of the conditional relative bias (ASDCRB) is appreciable; however, it is smallest for the time series methods. As expected, the expansion estimator g_1 does well with respect to the unconditional bias measure, AARB, but its conditional performance (ASDCRB) is quite poor.

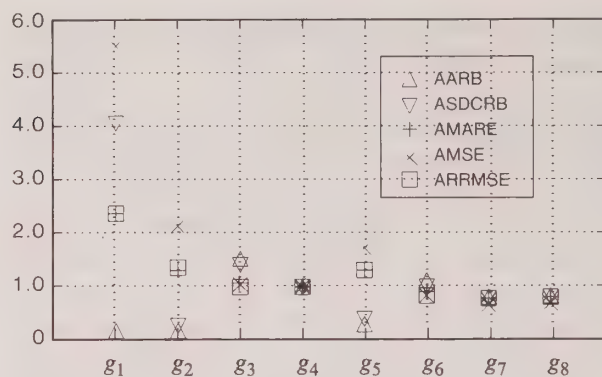


Figure 1. Evaluation Measures Relative to Fay-Herriot

Note: Relative ASDCRB for g_1 (= 18.98) not shown.

Table 2
Average Evaluation Measures

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
AARB	.001	.007	.097	.065	.018	.070	.053	.053
ASDCRB	.282	.016	.016	.015	.023	.010	.010	.010
AMARE	.269	.147	.115	.108	.136	.097	.087	.088
ARRMSE	.339	.192	.137	.137	.176	.120	.109	.111
AMSE (1,000's)	72,979	27,596	13,382	12,898	22,760	10,603	8,610	8,829

Figure 2 plots averages of $RRMSE_k$ for three size groups, namely small, medium and large small areas, based on the ranking of their true population totals at time T . They are divided up into these three groups because the relative errors of estimation would be expected to be larger for the smaller totals, and the plots do not contradict this expectation. Again, the time series methods g_7 and g_8 perform best. Note that the time series method g_6 , which assumes the small area effects to be independent over time, does not do as well. The unaveraged values of $RRMSE_k$ are given in Table 3. $RRMSE_9$ is relatively large because the total number of cattle and calves for area 9 is less than half that of any other small area. Areas 6 and 8 stand out within the medium size small areas as being most difficult to estimate by the smoothing methods. The reason for this is that, while there was an overall decline of about 16% in the total number of cattle and calves in the pseudo-population from June 1986 to January 1991, the decreases for areas 6 and 8 were the furthest from the average at 33% and 1%, respectively, so the ratio adjusted covariate would be least appropriate for those areas. Nevertheless, the time series methods g_7 and g_8 performed significantly better than the post-stratified estimator for areas 6 and 8. This is because the random walk model for the small area effects is able to track small areas which, like areas 6 and 8, progressively deviate from the model.

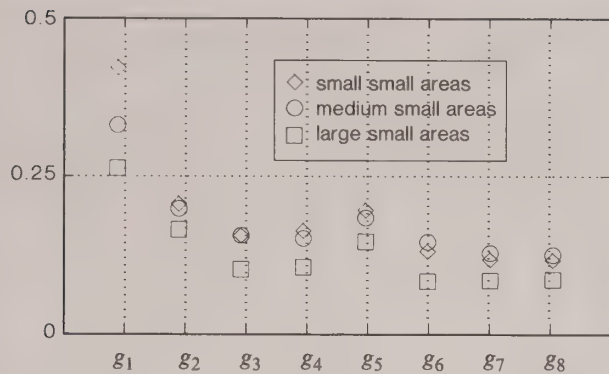


Figure 2. Relative Root Mean Squared Errors: Averaged within Size Groups

Table 3

Relative Root Mean Squared Errors and True Total Cattle and Calves for Small Areas

	Area	True Values	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Small Size	9	8,502	.580	.277	.342	.275	.277	.199	.160	.174
	10	18,990	.360	.196	.078	.113	.175	.097	.103	.104
	11	18,776	.339	.122	.122	.103	.112	.096	.086	.087
	12	19,819	.409	.237	.076	.152	.212	.123	.117	.117
	Average	16,522	.422	.208	.154	.161	.194	.129	.116	.120
Medium Size	1	27,595	.312	.206	.117	.130	.185	.120	.100	.102
	6	29,012	.306	.241	.256	.216	.224	.224	.168	.172
	7	23,600	.341	.121	.107	.094	.110	.088	.092	.092
	8	23,627	.383	.250	.155	.165	.219	.155	.146	.144
	Average	25,959	.336	.205	.159	.151	.185	.147	.126	.127
Large Size	2	35,592	.268	.171	.113	.110	.156	.096	.089	.088
	3	40,582	.241	.151	.087	.090	.137	.070	.072	.073
	4	42,396	.256	.160	.099	.103	.144	.080	.088	.089
	5	35,996	.270	.176	.091	.097	.160	.088	.085	.088
	Average	38,642	.259	.164	.098	.100	.149	.083	.083	.084

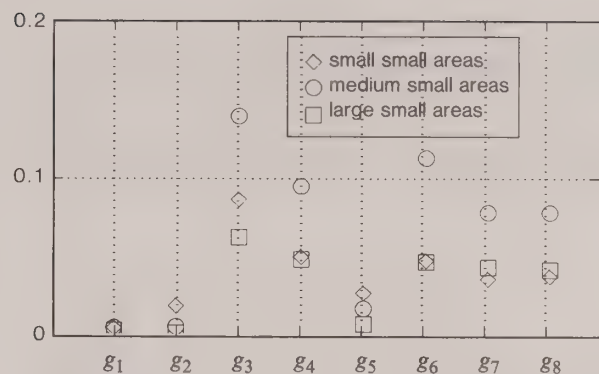


Figure 3. Absolute Relative Biases: Averaged within Size Groups

Table 4

Absolute Relative Biases and True Total Cattle and Calves for Small Areas

	Area	True Values	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
Small Size	9	8,502	.002	.047	.232	.139	.085	.099	.061	.069
	10	18,990	.002	.002	.006	.007	.003	.015	.026	.025
	11	18,776	.002	.009	.090	.052	.021	.062	.039	.037
	12	19,819	.000	.007	.019	.011	.007	.023	.024	.023
	Average	16,522	.001	.016	.087	.052	.029	.050	.037	.039
Medium Size	1	27,595	.001	.003	.093	.063	.007	.078	.044	.045
	6	29,012	.000	.001	.239	.157	.023	.195	.120	.123
	7	23,600	.000	.005	.088	.053	.014	.058	.062	.061
	8	23,627	.002	.008	.143	.106	.024	.124	.093	.091
	Average	25,959	.001	.004	.141	.095	.017	.114	.080	.080
Large Size	2	35,592	.000	.000	.095	.071	.009	.068	.049	.047
	3	40,582	.000	.001	.047	.041	.005	.029	.026	.025
	4	42,396	.001	.002	.066	.056	.008	.044	.057	.056
	5	35,996	.000	.000	.045	.029	.005	.048	.035	.039
	Average	38,642	.000	.001	.063	.049	.006	.047	.042	.042

Figure 3 and Table 4 are identical to Figure 2 and Table 3 in format, but show relative biases instead of relative root mean squared errors. The biases for both the expansion estimator g_1 and the post-stratified g_2 are negligible. For the smoothing methods the average absolute relative biases for medium size small areas are relatively large, mainly because of areas 6 and 8 for which the covariate is least appropriate. Among smoothing methods, the sample size dependent g_5 has the least bias because it is usually very close to the direct g_2 ; however, it also gains very little over g_2 with respect to mean squared error. Of the remaining smoothing methods the time series estimators g_7 and g_8 , which had the smallest mean squared error, also have the smallest bias. Nevertheless, the relative bias of these methods can be quite large, as in areas 6 and 8. In practice it would not be possible to estimate these biases; however, the possible size of the bias could be assessed using simulated sampling from a variety of plausible populations.

5. CONCLUDING REMARKS

It was seen by means of a simulation study that small area estimation methods obtained by combining both cross-sectional and time series data can perform better than those based only on cross-sectional data, with respect to both bias and mean squared error. However, the cost in terms of bias could still be substantial. A question of obvious importance is whether it is possible in practical situations to judge if the gains from any type of smoothing would outweigh the costs, and how to make this judgement.

The models for the simulation study were chosen on general considerations. However, in practice, suitable diagnostics similar to those employed in Pfeffermann and Barnard (1991) should be developed for survey data before any model-assisted method can be recommended. It should also be noted that the small area estimators could be modified to make them robust to mis-specification of the

underlying model as suggested by Pfeiffermann and Burck (1990), see also Mantel, Singh and Bureau (1993). Finally, modification and further extension of the methods presented in this paper to the more realistic case of correlated sampling errors should be investigated in the future.

ACKNOWLEDGEMENT

We would like to thank Jon Rao, Danny Pfeiffermann and M.P. Singh for useful discussions and comments on earlier versions of this paper. The comments and suggestions of an anonymous referee and an Associate Editor are also very much appreciated. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

APPENDIX

A.1 Variance Estimation for g_{2kt}

Let v_{kt} denote the conditional (given n_{hkt}) variance of g_{2kt} in (2.2). Then v_{kt} is given by (whenever $n_{hkt} > 0$ for all h at time t),

$$v_{kt} = \sum_h N_{hkt}^2 \left(n_{hkt}^{-1} - N_{hkt}^{-1} \right) \sigma_{hkt}^2, \quad (\text{A.1})$$

where σ_{hkt}^2 is the population variance for the intersection of the h -th stratum with the k -th small area at time t . The variance σ_{hkt}^2 can be estimated by the usual estimator s_{hkt}^2 for $n_{hkt} \geq 2$. Note that the estimate of the conditional variance v_{kt} also provides an estimate of the unconditional variance of g_{2kt} .

If $n_{hkt} = 1$, then we can use a synthetic value as an estimate of σ_{hkt}^2 which can be defined as $\sum (n_{hkt} - 1) s_{hkt}^2 / \sum (n_{hkt} - 1)$, the summation being over all k for which $n_{hkt} \geq 2$ within each (h, t) . If $n_{hkt} = 0$, v_{ht} of (A.1) is of course not defined. With the synthetic value of \bar{y}_{hkt} used in this case, we need a synthetic value of its mean squared error. For each (h, t) , it can be defined as

$$(\bar{X}_{hkt} / \bar{X}_{ht})^2 (n_{ht}^{-1} - N_{ht}^{-1}) s_{ht}^2 + (\widehat{\text{bias}})^2,$$

where $(\widehat{\text{bias}})^2$ will be taken as

$$\sum_{n_{hlt} > 0} ((\bar{X}_{hlt} / \bar{X}_{ht}) \bar{y}_{ht} - \bar{y}_{hlt})^2 / m_{ht},$$

where m_{ht} is the number of small areas with sample in stratum h at time t .

A.2 Estimation of Variance Components

Using the notation of (3.2), we here illustrate the method of moments for estimating variance components for the model of Section 3.1 in the special case when there is only one auxiliary variable X_{ht} , $Q_t = \tau^2 I$ and β_t follows a random walk, i.e., $G_t^{(1)} = I$. Let $F_t = (F_{1t}, \dots, F_{Kt})'$, $F_{kt} = (1, X_{kt})'$, $\beta_t = (\beta_{1t}, \beta_{2t})'$, and $B_t = \text{diag}(\gamma_1^2, \gamma_2^2)$. The parameter τ^2 is estimated by the solution of

$$\sum_{t=1}^T \sum_{k=1}^K (g_{2kt} - F_{kt}' \hat{\beta}_t)^2 / (v_{kt} + \tau^2) = T(K - 2).$$

If there is no positive solution, we set $\hat{\tau}^2 = 0$. Here $\hat{\beta}_t$ denotes the WLS estimate of β_t based on only the cross-sectional data at t . This is analogous to the method used in Fay and Herriot (1979) for cross-sectional data. An estimate of γ_i^2 can be obtained by solving (for $i = 1, 2$)

$$\sum_{t=2}^T (\hat{\beta}_{i,t} - \hat{\beta}_{i,t-1})^2 / (\gamma_i^2 + d_{ii}^{(t)}) = T - 1,$$

where $d_{ii}^{(t)}$ is the (i, i) -th element of $(F_{t-1}' U_{t-1}^{-1} F_{t-1})^{-1} + (F_t' U_t^{-1} F_t)^{-1}$.

REFERENCES

- BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error-components model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- BELL, W.R., and HILLMER, S.C. (1987). Time series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.
- BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- DUNCAN, D.B., and HORN, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association*, 67, 815-821.
- ERICKSEN, E.P. (1974). A regression method for estimating populations of local areas. *Journal of the American Statistical Association*, 69, 867-875.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21, U.S. Office of Management and Budget.
- GHOSH, M., and RAO, J.N.K. (1994). Small Area Estimation: an Appraisal. *Statistical Science*, 9, to appear.

- GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: University Press.
- JULIEN, C., and MARANDA, F. (1990). Sample design of the 1988 national farm survey. *Survey Methodology*, 16, 117-129.
- MANTEL, H.J., SINGH, A.C., and BUREAU, M. (1993). Benchmarking of small area estimators. *Proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, 920-925.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economics Statistics*, 9, 163-175.
- PFEFFERMANN, D., and BARNARD, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics*, 9, 73-84.
- PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.-E., and SINGH, M.P. Eds. (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley and Sons.
- RAO, J.N.K., and CHOUDHRY, G.H. (1993). Small area estimation: Overview and empirical study. Presented at the International Conference on Establishment Surveys, Buffalo, June 1993, to appear.
- RAO, J.N.K., and YU, M. (1992). Small Area Estimation by Combining Time Series and Cross-sectional Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.
- SALLAS, W.M., and HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.
- SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHAIBLE, W.L. (1978). Choosing weights for composite estimation for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.
- SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.
- SINGH, A.C., and MANTEL, H.J. (1991) State space composite estimation for small areas. *Proceedings: Symposium 91: Spatial Issues in Statistics*, Statistics Canada, Ottawa, November 1991, 17-25.
- TILLER, R. (1992). Time series modelling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.

Jackknife Variance Estimation of Imputed Survey Data

JOHN G. KOVAR and EDWARD J. CHEN¹

ABSTRACT

Imputation is a common technique employed by survey-taking organizations in order to address the problem of item nonresponse. While in most of the cases the resulting completed data sets provide good estimates of means and totals, the corresponding variances are often grossly underestimated. A number of methods to remedy this problem exists, but most of them depend on the sampling design and the imputation method. Recently, Rao (1992), and Rao and Shao (1992) have proposed a unified jackknife approach to variance estimation of imputed data sets. The present paper explores this technique empirically, using a real population of businesses, under a simple random sampling design and a uniform nonresponse mechanism. Extensions to stratified multistage sample designs are considered, and the performance of the proposed variance estimator under non-uniform response mechanisms is briefly investigated.

KEY WORDS: Item nonresponse; Hot deck imputation; Nearest neighbour imputation; Nonrandom nonresponse; Complex survey design.

1. INTRODUCTION

All sample surveys suffer from varying degrees of nonresponse. While total or unit nonresponse is often redressed by appropriate survey weight adjustment, most survey taking organizations resort to imputation in the case of item nonresponse. In this way, plausible values are inserted in place of missing or inconsistent entries, thus simplifying estimation of means and totals at all levels of aggregation. As early as the 1950's however, Hansen, Hurwitz and Madow (1953) recognized that treating the imputed values as observed values can lead to underestimation of variances of these estimators if standard formulae are used; underestimation which becomes more appreciable as the proportion of imputed items increases.

A number of remedies to overcome this problem have been advanced. In particular, Rubin (1987) proposed multiple imputation to estimate the variance due to imputation by replicating the process a number of times and estimating the between replicate variation. More recently, Särndal (1990) outlined a number of model assisted estimators of variance, while Rao and Shao (1992) proposed a technique that adjusts the imputed values to correct the usual or naive jackknife variance estimator. The Särndal, and Rao and Shao methods, are appealing in that only the imputed file (with the imputed fields flagged) is required for variance estimation. No auxiliary files are needed. Särndal's model assisted approach yields unbiased variance estimators, provided the model holds (Lee, Rancourt and Särndal 1991). The Rao and Shao adjusted jackknife method is design consistent as well as model unbiased (Rao 1992). But while the model assisted

approach requires different variance estimators for each imputation method, the adjusted jackknife method provides a unified approach that requires the implementation of only one estimator, the jackknife estimator, provided the imputed values are adjusted appropriately during the variance estimation stage.

In this paper we describe a simulation study that evaluates the adjusted jackknife variance estimator of Rao and Shao (1992). In Section 2 we motivate the present empirical study by demonstrating the characteristics of the naive variance estimator under four imputation methods in the case of simple random sampling. In Section 3 we briefly outline the Rao and Shao adjustment procedure and present the empirical results. Extensions to more complex designs and experiments with nonrandom nonresponse mechanisms are elaborated in Section 4. Finally, in Section 5 we offer some concluding remarks and recommendations, including areas for future study.

2. BACKGROUND

Following the notation of Rao (1992), we suppose that in a sample s , of size n , m units respond to item y , while $n - m$ units do not. Denote by y_i^* the imputed value for unit i , $i \in s - s_r$, where s_r is the set of responding units. The usual estimator of the mean \bar{Y} under simple random sampling, based on the imputed file is given by

$$\bar{y}_I = \frac{1}{n} \left(\sum_{i \in s_r} y_i + \sum_{i \in s - s_r} y_i^* \right). \quad (1)$$

¹ John G. Kovar, Business Survey Methods Division; Edward J. Chen, Social Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2.1 Imputation Methods

In the present simulation study we consider four simple methods of imputation, namely the mean of respondents, ratio, nearest neighbour and hot deck imputation methods. The reader is referred to Kalton and Kasprzyk (1986) for a thorough review of the topic of imputation. The simplest and most intuitive method of imputation, when the interest lies in estimating the mean of the item y , is to impute all missing items with the mean of the observed responding units. The imputed value y_i^* , for unit i , under the mean imputation method, is thus given by

$$y_i^* = \bar{y}_m = \sum_{j \in s_r} y_j / m. \quad (2)$$

In this case, the estimator of the population mean \bar{Y} in (1) reduces to the estimator $\bar{y}_I = \bar{y}_m$. Due to the fact that this method has the undesirable property of distorting the distributions, it is used in practice usually only as a last resort. It is included here for illustrative purposes.

Secondly, we consider a ratio imputation method based on the assumption that a correlated auxiliary variable x , is available, and that the ratio \bar{y}/\bar{x} is the same in the s_r and $s - s_r$ sets, as would be the case if the nonresponse occurred at random, for example. Under the ratio imputation method, we impute the predicted value in place of the missing y_i as follows:

$$y_i^* = \frac{\bar{y}_m}{\bar{x}_m} x_i, \quad (3)$$

where \bar{x}_m is the mean of the x values of the respondent set s_r . The estimator of the population mean \bar{Y} in (1) reduces to the double sampling estimator $\bar{y}_I = (\bar{y}_m/\bar{x}_m)\bar{x}$, by considering the respondents as the second phase sample.

The third imputation technique we consider is the nearest neighbour (NN) method. Under this method, the missing value is filled in by an observed value of another unit from the set s_r , whose distance to the nonresponding unit is minimum. In practice the distance functions used are usually the ζ_1 , ζ_2 , or ζ_∞ Minkowski's norms based on the auxiliary x -variables, assumed observed for all units in s . Thus

$$y_i^* = y_j, j \in s_r, \text{ such that } \|x_i - x_j\| \text{ is minimized, } (4)$$

where $\|\cdot\|$ is one of the above mentioned norms.

The above three methods are often labelled deterministic, since, given the sample of respondents, the imputed values are determined uniquely. The fourth imputation method considered in this study, the hot deck method (HD), is non-deterministic, since the imputed values are chosen at random from the respondent set. While in practice imputation classes are often created and

some sort of sequential procedure is usually implemented, we consider here the pure hot deck, whereby the donor unit (j) is chosen at random, with replacement, from the entire set s_r , that is,

$$y_i^* = y_j, j \in s_r. \quad (5)$$

2.2 Variance Due to Imputation

Treating the imputed values as observed values, leads to the incorrect variance estimator

$$v_{naive} = (1 - f)s_I^2/n, \quad (6)$$

where s_I^2 is the sample variance of the complete sample of responding and imputed values, and $(1 - f)$ is the finite population correction factor ($f = n/N$). It can be easily shown that the true variance of the estimator \bar{y}_I in (1), $V(\bar{y}_I)$, can be written as (Särndal 1990)

$$V(\bar{y}_I) = V_{sam} + V_{imp} + V_{mix}, \quad (7)$$

where V_{sam} is the sampling variance component, V_{imp} is the variance introduced by the imputation method in question and V_{mix} is a covariance term between V_{sam} and V_{imp} which in most cases is negligible or zero. An estimator of V_{sam} could be obtained by adding to v_{naive} a term to correct for the fact that the standard formula understates the sampling variance component when there are imputed values in the data set. To estimate $V(\bar{y}_I)$, however, an additional component of variance due to the imputation mechanism, V_{imp} , must be estimated. This may be done explicitly, as in Rubin's (1987) multiple imputation, or by modifying common variance formulae as in Särndal (1990) and Rao and Shao (1992). Note that the interest lies in estimating the variance of the estimator at hand, that is, $V(\bar{y}_I)$, not the variance of an estimator that would have been obtained had there been no nonresponse.

2.3 Variance Underestimation

To illustrate the seriousness of the underestimation of $V(\bar{y}_I)$ by v_{naive} , and the dependence of the degree of underestimation on the imputation method, we first describe the simulation study used for this purpose. We consider a data set of 5,620 units with two variables: An auxiliary variable x , the Gross Business Income, available for all units, that can be used as a measure of size, and a related purchase variable y . The correlation between x and y in this particular data set is of the order of 0.92. Simple random samples of size 200 were selected without replacement. A fixed proportion of units were identified at random as nonrespondents, having their y -values deleted and imputed according to one of the four methods described above. Various rates of nonresponse were generated, though, for the most part, we confine our reporting to results based on 5 and 30% nonresponse rates.

To evaluate the performance of the proposed variance estimators, we calculate the percent relative bias of the variance estimator v_{\cdot} , given by

$$\text{Rel.Bias}(v_{\cdot}) = \sum_{k=1}^K \frac{(v_k - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100, \quad (8)$$

where $V(\bar{y}_I)$ is obtained through simulation, and v_k is the k -th realization of the K simulated variance estimates in question. Similarly, the percent relative stability of the variance estimators is given by

$$\text{Rel.Stab.}(v_{\cdot}) = \sum_{k=1}^K \frac{\sqrt{(v_k - V(\bar{y}_I))^2/K}}{V(\bar{y}_I)} \times 100. \quad (9)$$

All simulations were performed on an IBM PC, using Microsoft's Fortran 77, Version 5.0. In the case of simple random sampling, results are based on averages of 100,000 replications ($K = 100,000$). With this number of replicates, the reported relative bias values were observed not to vary by more than one percentage point. The results are summarized below in Table 1 for the case of 5 and 30% nonresponse rates.

Table 1

Underestimation of Variance of \bar{y}_I by the Naive Estimator
Under Four Imputation Methods, and 5 and 30%
Nonresponse Rates

Non-response Rate	Variance Estimator	Imputation Method			
		Mean	HD	Ratio	NN
5%	$V(\bar{y}_I)$	9.9	10.3	9.5	9.5
	v_{naive}	8.9	9.4	9.2	9.3
	Rel.Bias (v_{naive})	-10.7%	-9.4%	-2.5%	-2.2%
30%	$V(\bar{y}_I)$	13.5	16.5	10.1	10.3
	v_{naive}	6.5	9.4	8.5	9.0
	Rel.Bias (v_{naive})	-51.4%	-43.4%	-15.3%	-12.8%

First, we note in Table 1, that the naive estimator underestimates the true variance of \bar{y}_I by 10.7% in the case of mean imputation at a 5% level of nonresponse. About half of this underestimation is due to the fact that v_{naive} underestimates V_{sam} and the other half is due to the fact that v_{naive} ignores the component V_{imp} . Särndal (1990) obtains very similar results with respect to the partitioning of the underestimation in the case of mean imputation. Secondly, in the first row of Table 1, the true variance of \bar{y}_I is larger in the case of the hot deck imputation as compared to the mean imputation, due to the procedure's inherent variability (*i.e.*, the V_{imp} component is larger). By contrast, $V(\bar{y}_I)$ is slightly lower in the case of the ratio and nearest

neighbour imputation methods, since V_{imp} decreases as the imputation procedure is better able to predict the true unobserved values (Särndal 1990), as is the case in the present study due to the relatively high correlation between the x and y variables. Thirdly, as can be seen in Table 1, $V(\bar{y}_I)$ increases while v_{naive} decreases as the nonresponse rate becomes more elevated. As such, the underestimation of $V(\bar{y}_I)$, when the imputed values are treated as observed values, becomes more serious as the proportion of missing items increases. The problem is more pronounced in the case of the mean and hot deck imputation methods, which do not use auxiliary information. Note that underestimation of variance in the order of 50%, as was observed in this case, can lead to confidence intervals that are about 30% too short and to declaration of significance when none exists. Also of note is the similar behaviour of the ratio and nearest neighbour methods which will be exploited later.

3. JACKKNIFE VARIANCE ESTIMATOR

Let $\bar{y}_I(j)$ be the imputed estimator of \bar{Y} obtained when the j -th unit is deleted from the sample. Then, in the case of simple random sampling, a naive jackknife variance estimator of \bar{y}_I is given by

$$\bar{v}_J = \frac{n-1}{n} \sum_{j=1}^n [\bar{y}_I(j) - \bar{y}_I]^2, \quad (10)$$

which can be shown to reduce to v_{naive} (Rao 1992).

3.1 Imputed Value Adjustment

In order to produce the "correct" (Rao 1990) jackknife variance estimator, Rao (1992) proposed to adjust the imputed values as described below. Intuitively, the adjustment is necessary whenever a responding unit is deleted from a jackknife replicate, since in the case of most imputation methods, all the imputed values depend directly or indirectly on the observed value that was deleted. This is clear in the case of mean imputation and ratio imputation, where all respondents contribute directly to the mean \bar{y}_m , but is less evident in nearest neighbour and hot deck imputation methods where the deleted unit contributes to the imputation process only in the sense that it is not available to be selected as a donor. Thus, whenever a responding unit is deleted, *all* imputed values in the sample must be adjusted before the "delete-one" imputed estimator of the mean is computed. The adjustment must clearly be a function of the imputation method used. In the case of the mean and the hot deck imputation methods, it can be shown that the following adjustment is appropriate (Rao 1992; Rao and Shao 1992). Let $z_i^*(j)$ be the adjusted value of the i -th imputed unit y_i^* , when the j -th unit has been deleted. Then $z_i^*(j)$ is given by

$$z_i^*(j) = \begin{cases} y_i^* + [\bar{y}_m(j) - \bar{y}_m] & \text{if } j \in S_r \\ y_i^* & \text{if } j \in S-S_r. \end{cases} \quad (11)$$

In other words, no adjustment is necessary if the deleted unit (j), has itself been imputed; that is, unit j is a non-respondent. In the case of the mean imputation, for example, when $j \in S_r$, the adjusted value reduces to $\bar{y}_m(j)$, the mean of the remaining $m - 1$ respondents, as desired.

The jackknife variance estimator is evaluated by first computing the adjusted imputed estimator $\bar{y}_I^q(j)$, as

$$\bar{y}_I^q(j) = \sum_{\substack{i \in S \\ i \neq j}} z_i^*(j) / (n - 1), \quad (12)$$

and then letting

$$v_J(\bar{y}_I) = \frac{n-1}{n} \sum_{j=1}^n [\bar{y}_I^q(j) - \bar{y}_I]^2. \quad (13)$$

It can be shown that the adjusted jackknife variance estimator reduces to the correct variance estimator in the case of the mean imputation (Rao 1990), and provides a consistent estimator in the case of the hot deck imputation (Rao and Shao 1992).

In the case of the ratio imputation, the adjusted values are given by

$$z_i^*(j) = \begin{cases} y_i^* + \left[\frac{\bar{y}_m(j)}{\bar{x}_m(j)} x_i - \frac{\bar{y}_m}{\bar{x}_m} x_i \right] & \text{if } j \in S_r \\ y_i^* & \text{if } j \in S-S_r, \end{cases} \quad (14)$$

where $\bar{x}_m(j)$ is the mean of the $m - 1$ sample values of x of the responding units when unit j is deleted. The jackknife variance estimator $v_J(\bar{y}_I)$ is then computed as in (13) above, yielding the correct variance estimator. Furthermore, Rao (1992) shows that not only is the adjusted jackknife variance estimator design consistent (p -consistent) under uniform nonresponse irrespective of the model, but is also design-model unbiased (pm -unbiased) under the model (15) and any nonresponse mechanism that does not depend on the y -values.

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i, \\ \text{cov}_m(y_i, y_j) = 0 \quad i \neq j \in S. \quad (15)$$

Since the naive variance estimator under the nearest neighbour imputation was observed to behave much like the naive variance estimator under the ratio imputation, the adjustment for the ratio imputation given in (14) was used in the case of the nearest neighbour imputation. As well, an alternate adjustment was considered, whereby unit i was re-imputed using the nearest neighbour method,

whenever the deleted unit (j) was used to impute unit i . That is, adjustment takes place only if the deleted unit is a respondent (as above), but only those nonrespondents in the j -th jackknife replicate that were actually imputed using unit j are re-imputed by one of the $m - 1$ remaining donors. (This corresponds to imputing the second nearest neighbour for these units.) We note that no theoretical justification exists for either of these adjustments. Since the latter adjustment performed worse than the ratio adjustment in our examples, and since its eventual implementation in production would be cumbersome, we omitted it from further consideration, even though it was always observed to be conservative.

We would like to stress here that for all imputation methods the adjustments are only performed for the purpose of variance estimation and can be made temporarily while the variance estimation program executes. No permanent adjustments are required on the imputed file used for the estimation of means and totals, though the imputed fields must be flagged appropriately.

3.2 Empirical Results

The jackknife variance estimator with adjustments corresponding to the four imputation methods described above, was computed in addition to v_{naive} in the simulation study outlined in Section 2. Nonresponse rates of 5 and 30% were considered and the relative biases were calculated. They are summarized in Table 2 below.

Table 2
Relative Biases of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under 5 and 30% Nonresponse Rates

Non-response Rate	Variance Estimator	Imputation Method			
		Mean	HD	Ratio	NN
in percent					
5%	v_{naive}	-10.7	-9.4	-2.5	-2.2
	v_J	2.7	3.6	3.4	3.7
30%	v_{naive}	-51.4	-43.4	-15.3	-12.8
	v_J	3.3	1.9	3.0	5.3

Since the adjusted jackknife variance estimator is design consistent (p -consistent) (Rao 1992), it performs well in the case of the mean, hot deck and ratio imputation under uniform response mechanism, as expected. (Equally good performance was observed with other data sets which do not follow the model (15) as well, but more work is needed on this front.) Of note is the relatively good performance under the nearest neighbour imputation. The proposed estimator tends to be somewhat conservative, due, in small part, to the fact that it does not incorporate the finite population correction.

4. EXTENSIONS

While the adjusted jackknife variance estimator has been shown to perform well in the case of simple random sampling under uniform nonresponse mechanism in one imputation class, we consider here extensions to more complex design, to more than one imputation class, and to nonrandom response mechanisms.

4.1 Complex Designs

In this section we describe a simulation study that evaluates the Rao and Shao (1992) adjusted jackknife variance estimator in comparison to the naive variance estimator, in the case of stratified multistage sampling and hot deck imputation. In particular, data from the Canadian Survey of Consumer Finances (SCF) that follows the design of the Canadian Labour Force Survey will be used. The variable of interest, y , is the total household income. The SCF follows a complex stratified multistage design with the primary sampling units (psu's) in the strata used in this study selected with probability proportional to the number of dwellings. Generally speaking, the psu's are collections of dwellings, corresponding to city blocks in urban areas and to groups of Census Enumeration Areas (EA's) in rural regions. We used as a population a sample of 3,870 households in 30 strata and sampled two psu's in each stratum. As in the case of the simple random sampling study, 5 and 30% uniform nonresponse rates were generated at the household level. The missing values were then imputed using the hot deck imputation method described in Rao and Shao (1992). Briefly, the imputation method consists of selecting the donors from the respondent set with replacement, with probability proportional to the survey weight of the donors.

We first consider the case of a single imputation class. Let y_{hik} be the observed value for the k -th unit in the i -th psu and the h -th stratum ($k = 1, \dots, n_{hi}, i = 1, \dots, n_h, h = 1, \dots, L, n = \sum \sum n_{hi}$), and let y_{hik}^* be the corresponding imputed value whenever the (hik) unit is a nonrespondent, that is, whenever $(hik) \in s-s_r$. The imputed estimator of Y is then given by

$$\hat{Y}_I = \sum_{(hik) \in s_r} w_{hik} y_{hik} + \sum_{(hik) \in s-s_r} w_{hik} y_{hik}^*, \quad (16)$$

where w_{hik} is the survey weight corresponding to unit (hik) . Under the above hot deck imputation scheme, \hat{Y}_I is asymptotically unbiased (Rao and Shao 1992).

The expectation of \hat{Y}_I under the hot deck imputation procedure can be written as (Rao and Shao 1992):

$$\begin{aligned} E_*(\hat{Y}_I) &= \left[\sum_{(hik) \in s_r} w_{hik} y_{hik} / \sum_{(hik) \in s_r} w_{hik} \right] \times \sum_{(hik) \in s} w_{hik} \\ &= [\hat{S}/\hat{T}] \times \hat{U}, \end{aligned} \quad (17)$$

thus defining the terms \hat{S} , \hat{T} and \hat{U} . The jackknife "delete-one" values are then given by

$$\hat{S}(gj) = \sum_{\substack{(hik) \in s_r \\ h \neq g}} w_{hik} y_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik) \in s_r \\ i \neq j}} w_{gik} y_{gik}, \quad (18)$$

$$\hat{T}(gj) = \sum_{\substack{(hik) \in s_r \\ h \neq g}} w_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik) \in s_r \\ i \neq j}} w_{gik},$$

whenever the j -th psu in the g -th stratum is deleted. The adjustment of the imputed values is performed whenever the (gj) -th psu is deleted, $(hi) \neq (gj)$, and $(hik) \in s-s_r$, by letting

$$z_{hik}^{(gj)} = y_{hik}^* + \left[\frac{\hat{S}(gj)}{\hat{T}(gj)} - \frac{\hat{S}}{\hat{T}} \right]. \quad (19)$$

Then, analogous to (12) and (13), the jackknife variance estimator is evaluated by first computing the adjusted imputed estimator \hat{Y}_I^a when the (gj) -th psu is deleted as

$$\begin{aligned} \hat{Y}_I^a(gj) &= \hat{S}(gj) + \sum_{(hik) \in s-s_r} w_{hik} z_{hik}^{(gj)} \\ &\quad + \frac{n_g}{n_g - 1} \sum_{\substack{(hik) \in s-s_r \\ i \neq j}} w_{gik} z_{gik}^{(gj)}, \end{aligned} \quad (20)$$

and then setting

$$v_J(\hat{Y}_I) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_I^a(gj) - \hat{Y}_I)^2. \quad (21)$$

It can be shown that v_J as defined in (21), is a consistent estimator of the variance of \hat{Y}_I (Rao and Shao 1992).

We generated 10,000 samples of 60 psu's selected with probability proportional to size, and subjected the selected households to 5 and 30% uniform nonresponse. We then computed the naive variance estimator, and the adjusted jackknife variance estimator, v_J , in (21). The relative bias (8) and the relative stability (9) were computed for both of the variance estimators, and are summarized in Table 3 below.

Table 3

Relative Bias and Relative Stability (in Parentheses) of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under 5 and 30% Nonresponse, in the Case of Stratified Multistage Sampling

Variance Estimator	Nonresponse Rate	
	5%	30%
	in percent	
v_{naive}	-10.3 (88)	-43.7 (84)
v_J	-0.9 (97)	1.2 (124)

As can be seen in Table 3, the naive variance estimator underestimates the true variance of Y at rates comparable to the simple random sampling case (Table 2), with the underestimation becoming more serious as the nonresponse rate increases. The adjusted jackknife variance estimator, on the other hand, performs well at both levels of nonresponse, at a relatively modest cost of a slight decrease in the stability of the variance estimator, as compared to v_{naive} .

4.2 Imputation Classes

Under the same sample design as in Section 4.1, we also considered the case of more than one imputation class as is the case in practice. The household size, known for all households in the sample, was used to form two imputation classes, namely one member households and more than one member households. This was done under the assumption that the propensity to respond is different between these two classes, while uniform response probability was assumed within the imputation classes. Two nonresponse schemes were evaluated. The first assumes a 5% uniform nonresponse in the single member household class and 10% uniform nonresponse in the multiple member household class, while the second scheme assumes rates of 25 and 30% in each of the classes respectively. The hot deck imputation, the imputed value adjustments, and the adjusted total calculations in (20), $\hat{Y}_v^a(gj)$, were performed independently within each imputation class denoted by v . The terms $\hat{Y}_v^a(gj)$ were then summed over the two imputation classes, yielding $\hat{Y}_J^a(gj)$, which was used in (21) to provide the estimate v_J . The results are summarized in Table 4.

Table 4

Relative Bias and Relative Stability (in Parentheses) of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under Two Nonresponse Schemes, in the Case of Stratified Multistage Sampling and Two Imputation Classes

Variance Estimator	Nonresponse Rate	
	5% and 10%	25% and 30%
	in percent	
v_{naive}	-16.7 (87)	-40.2 (84)
v_J	-1.0 (103)	1.1 (127)

As can be seen in Table 4, the adjusted jackknife variance estimator v_J , performs well under both nonresponse schemes. The results, along with those in Table 3, demonstrate the consistency and the reasonably good stability of the adjusted jackknife variance estimator, even in cases of elevated nonresponse rates.

4.3 Nonrandom Nonresponse

As demonstrated above, the adjusted jackknife variance estimator performs well when the nonresponse is random within imputation classes. To study its robustness against the uniform response mechanism assumption, we use the data set described in Section 2, and generated nonresponse as outlined in Lee, Rancourt and Särndal (1991). In particular, the probability of nonresponse is assumed to be related to the x -variable in two distinct ways:

$$P_L = 1 - \exp(-c_L x), \quad (22)$$

$$P_S = \exp(-c_S x), \quad (23)$$

where the constants c_L and c_S are chosen such that an expected 30% nonresponse rate is achieved. In the model P_L given in (22) the nonresponse is positively correlated with the x -variable, implying that large (L) units are more likely not to respond. The opposite is true in the model P_S given in (23), under which smaller (S) units are more likely not to respond. Imputation methods which ignore the x -variable (mean and hot deck) are expected to yield estimators of \bar{Y} that underestimate the true mean under nonresponse model (22) and over estimate the true mean under the model (23). However, imputation methods that incorporate the auxiliary variable into the procedure (ratio and nearest neighbour), can be expected to produce better estimates of the mean. This has been confirmed by simulation as shown in Table 5 below. As before, 100,000 replicates were used.

Table 5

Estimates of the Mean \bar{Y} as Percent of the True Mean when the Nonresponse is not Random, and the Nonresponse Rate is an Expected 30%

Nonresponse Model	Imputation Method			
	Mean	HD	Ratio	NN
	in percent			
P_L	60.4	60.4	94.7	93.5
P_S	132.7	132.7	102.0	101.4

Clearly, variance estimation is of no interest when the point estimators themselves are highly biased as is the case for the mean and hot deck methods. However, in the case of the ratio and nearest neighbour methods, under which the point estimators perform better, we investigated the performance of the adjusted jackknife variance estimator, as well as an estimator proposed by Särndal (1990), which can be written as (Rao 1992):

$$\begin{aligned} v_s(\bar{y}_I) = & \left(\frac{\bar{x}}{\bar{x}_m}\right)^2 \frac{1}{m(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2 \\ & + \left(\frac{\bar{y}_m}{\bar{x}_m}\right) \frac{2m}{n^2(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m}\right) x_i \quad (24) \\ & + \left(\frac{\bar{y}_m}{\bar{x}_m}\right)^2 \frac{1}{n(n-1)} \sum_{i \in s} (x_i - \bar{x})^2, \end{aligned}$$

provided that the finite population correction factor is ignored, and that $(n - 1)/n \cong 1$ and $(m - 1)/m \cong 1$. The results are summarized in Table 6.

Table 6

Relative Bias of the Naive Variance Estimator, the Adjusted Jackknife Variance Estimator and Särndal's Variance Estimator Under 30% Nonrandom Nonresponse

Nonresponse Model	Variance Estimator	Imputation Method	
		Ratio	NN
in percent			
P_L	v_{naive}	-22.7	-54.6
	v_J	3.9	-37.5
	v_S	-2.6	-36.8
P_S	v_{naive}	-4.0	-0.7
	v_J	3.7	7.2
	v_S	2.8	4.5

In the case of the ratio imputation, the naive variance estimator performs quite differently under the two non-response models (-22.7 versus -4.0%). This is due to the fact that while the reduction in effective sample size tends to decrease the variance in both cases, under the P_L model disproportionately more large units are missing which tends to accentuate this effect, whereas under the P_S model, where disproportionately more small units are missing, this effect tends to be partly compensated for. Secondly, the adjusted jackknife variance estimator performs well in the case of ratio imputation, but relatively poorly in the case of nearest neighbour imputation. This is due to the fact that the present data set follows the usual linear model (15) fairly well and the adjusted jackknife variance estimator has been shown to be model unbiased (Rao 1992) in the case of the ratio imputation. On the other hand, the ratio adjustment does not work well in the case of nearest neighbour imputation when the nonresponse is not uniform. The alternate adjustment for the nearest neighbour imputation described in Section 3, performs equally poorly in absolute terms (not shown here), though the estimates are always conservative. Thirdly, the performance of

Särndal's estimator, v_S , is roughly equivalent to that of the adjusted jackknife estimator under either the ratio or the nearest neighbour imputation methods, and non-random nonresponse that depends only on x .

In cases where the response mechanism is not random, and when the propensity to respond is related to the variable subject to nonresponse (y), the point estimators are themselves severely biased under all four imputation methods. As such, variance estimation is of little interest, as the real interest lies in estimating the mean squared error. That is, more attention needs to be concentrated on improving the point estimates and their bias. Some preliminary results on this front have been put forth by Rancourt, Lee and Särndal (1992).

5. CONCLUDING REMARKS

It is well known that the usual variance estimator understates the variance of the estimate of \bar{Y} in the presence of imputed values if these values are treated as having been observed. In this study we again demonstrated the high degree of underestimation of the naive variance estimator in the presence of imputed data. Several imputation methods were considered in order to illuminate the dependence of the degree of underestimation on the method of imputation. We evaluated a unified jackknife variance estimator proposed by Rao and Shao (1992), an estimator that incorporates the variance due to imputation component. The study demonstrated some desirable properties of the proposed estimator in the case of both simple random sampling as well as complex survey designs. Our findings can be summarized as follows.

- (1) The extent of variance underestimation is highly dependent on both the imputation method's ability to predict the true values, and its ability to preserve the natural variation in the data.
- (2) The proposed adjusted jackknife variance estimator offers a unified approach to variance estimation of imputed data, that is easy to implement under a number of imputation methods and under designs of varying complexity.
- (3) Operationally, no modifications to the original imputed file are necessary and the estimation of means and totals is thus unaffected by the need to estimate variances.
- (4) The proposed method is easily extended to more complex designs, more than one imputation class and, with care, to the case of nonrandom nonresponse that depends only on available auxiliary variables.
- (5) The adjusted jackknife variance estimator performs well whenever the nonresponse is uniform or the usual linear model holds, demonstrating the fact that the estimator is both design consistent as well as design-model unbiased.

- (6) In the case of the P_L model, under which units with large y -values are more likely to not respond, all three variance estimators perform extremely poorly.
- (7) In the case of y -dependent nonresponse, better imputation techniques are needed and the bias of the point estimators needs to be studied further. Here the issue is primarily that of estimating the mean square error rather than the variance.

Given the relatively high degree of imputation in today's surveys, at least within some imputation classes, it is clear that the effect of imputation on variance estimation cannot be ignored. An overestimation of precision can lead to confidence intervals that are too short and to spurious declaration of significance. If implementation of the above suggested methods is deemed too onerous in any particular circumstance, at the very least studies should be conducted to evaluate the impact of imputation in some representative cases. An *ad hoc* variance inflation factor could then be implemented. With the emergence of generalized estimation software, however, there seems to remain little reason for not implementing variance estimators which correctly account for the effect of imputation.

There clearly remain many unsolved, and perhaps unsolvable problems. To begin with, much more theoretical work is needed with respect to nearest neighbour imputation. The jackknife adjustments considered for this imputation method fail to perform as well as those applied to the other methods. Perhaps smoother alternatives to the nearest neighbour method need to be developed. Secondly, the robustness of the proposed estimator must be investigated. It is clear that satisfactory performance can be obtained if the model (15) holds, and when nonresponse is random. Limited failure of either one of these conditions did not seem to detract from the good performance of the jackknife estimator in our limited experience, but further research along these lines is warranted. Departures from both of the conditions simultaneously are yet to be investigated. Cases of nonrandom nonresponse when the propensity of nonresponse is related to the y -variable are even less well understood, though the emphasis in this case must be placed on the estimation of the mean square error rather than the variance. Thirdly, comparisons to multiple imputation results should be considered. It must be recognized, however, that proper imputation methods (Rubin 1987) must first be established. We note that none of the imputation methods studied within are proper with respect to multiple imputation.

Extensions to other imputation methods and other parameters of interest should be undertaken. This study was limited to four simple imputation methods. In practice, much more complicated methods are used, often in conjunction with each other. The impact of more than one imputation method on the estimation of variance has

been studied by Rancourt, Lee and Särndal (1993); more work is needed. With respect to other, more complicated methods of imputation, the effect of adding theoretical residuals to imputed data can, for example, be considered. However, this technique only addresses the underestimation of V_{sam} by v_{naive} and ignores the effect of V_{imp} . Finally, other parameters, such as the median for example, and the effect of imputation on their variance are yet to be evaluated. Multivariate extensions can likewise be considered: estimation of correlations, ratios and regression parameters in the presence of imputation would likely be of interest.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. J.N.K. Rao for his continuous encouragement and support, and the Associate Editor for his constructive comments.

REFERENCES

- HANSEN, M., HURWITZ, W., and MADOW, W. (1953). *Sample Survey Methods and Theory*. (Volume 2), New York: J. Wiley, 139-141.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1991). Experiments with variance estimation from survey data with imputed values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 690-695.
- RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Unpublished report, Statistics Canada.
- RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Unpublished report, Statistics Canada.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1992). Bias corrections for survey estimates from data with imputed values for nonignorable nonresponse. *Proceedings 1992 Annual Research Conference, Bureau of the Census*, 523-539.
- RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 374-379.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.
- SÄRNDAL, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. Special invited lecture. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality, Statistics Canada*, 337-350.

Estimation in Overlapping Clusters with Unknown Population Size

D.S. TRACY and S.S. OSAHAN¹

ABSTRACT

Two sampling strategies for estimation of population mean in overlapping clusters with known population size have been proposed by Singh (1988). In this paper, ratio estimators under these two strategies are studied assuming the actual population size to be unknown, which is the more realistic situation in sample surveys. The sampling efficiencies of the two strategies are compared and a numerical illustration is provided.

KEY WORDS: Overlapping clusters; Clustering before sampling; Mean square error; Relative efficiency.

1. INTRODUCTION

In cluster sampling, clusters are formed either before selecting the sample (CBS) or after selecting the sample (CAS). In both cases, clusters may be overlapping or non-overlapping. For non-overlapping clusters, much work by several researchers is available in the literature. However, there are many practical sampling situations where one gets overlapping clusters. For example, overlapping clusters may exist in some regional epidemiological survey for a contagious disease like mycobacterim tuberculosis (T.B.), becoming very prevalent with the spread of AIDS (Gifford-Jones 1993). Clusters here may be formed around infected individuals or closely associated individuals who are more vulnerable to the same type of infection. A similar situation may exist in an ecological survey where clusters are formed around the factories burning coal and emitting polyaromatic hydrocarbons (PAH's) which are potent cancer causing compounds. Clusters are formed on the basis of the intensity of such gases, and surveys may be required in order to control air pollution which causes lung diseases like bronchitis. For overlapping clusters, one can refer to the limited work done by Goel and Singh (1977), Agarwal and Singh (1982) and Amdekar (1985). But the methodologies developed by them suffer from one limitation or the other.

Recently, Singh (1988) has developed a very simple estimator for a population mean using two sampling strategies in the CBS system assuming known population size. In the first strategy, clusters are selected with equal probabilities, whereas in the second case selection probabilities are taken proportional to cluster size. The elements within the clusters are selected with equal probability in both the cases. But it is unrealistic to assume that the actual population size is known. If it is the case, then all the duplicates in the population are known *a priori*, and one

could easily remove them to increase the efficiency of the sampling design. Hence, the estimators of the population mean studied by Singh (1988) need an improvement in order to be practicable, as they depend on the actual population size. This limitation in the methodology has motivated the present work.

We propose two sampling strategies in the CBS system with simple ratio estimators for the population mean, which do not depend on the actual population size. As in Singh (1988), an equal probability with replacement sampling scheme is used for selecting the clusters in the first strategy, whereas in the second, an unequal probability sampling scheme is used. The elements within the clusters are selected with an equal probability without replacement sampling scheme in both strategies.

The population of N units under consideration is expressible in the form of K overlapping clusters with N_i units in the i -th cluster and $\sum_{i=1}^K N_i = M \geq N$, the unknown actual population size, (equality holds only for non-overlapping clusters). A population unit may be included in more than one cluster. Let y be the characteristic of interest and let the population mean be \bar{Y} .

Define

$$Z_{ij} = Y_{ij}/F_{ij}, \quad W_{ij} = 1/F_{ij}; \quad i = 1, 2, \dots, K,$$

$$\text{and } j = 1, 2, \dots, N_i$$

where Y_{ij} is the value of y for the j -th unit in the i -th cluster and F_{ij} its frequency of occurring in K clusters.

When clusterwise data on units are available on the computer, the values of these frequencies for overlapping clusters may be easily available. As for the example considered earlier in epidemiology, suppose we have data available for households or individuals along with their identification labels like house numbers or social insurance

¹ D.S. Tracy and S.S. Osahan, Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario N9B 3P4.

numbers/health card numbers on the computer. Then, by giving a simple command to the computer, a researcher can easily extract information about the repetition of a certain unit from its label in different clusters. Also, in case we have a map of the overlapping clusters and the criterion for forming clusters does not allow the elimination of duplicacy of units in the different clusters, the values of such frequencies may be known.

The two strategies are discussed in section 2 and their efficiencies are compared in section 3.

2. THE TWO STRATEGIES

The two proposed strategies are discussed in Sections 2.1 and 2.2. Their comparison is undertaken in Section 3.

2.1 Strategy A

This strategy consists of the following steps:

- Select k clusters out of K by simple random sampling with replacement (SRSWR).
- From the i -th selected cluster of size N_i ($i = 1, \dots, K$), select n_i elementary units by simple random sampling without replacement (SRSWOR).

Theorem 1. The ratio estimator under SRS

$$\bar{z}_{RS} = \hat{Y}_{RS}/\hat{N}_{RS} = \frac{K}{k} \sum_{i=1}^k N_i \bar{z}_i \Big/ \frac{K}{k} \sum_{i=1}^k N_i \bar{w}_i \quad (1)$$

has relative bias, to the first order of approximation,

$$RB(\bar{z}_{RS}) \doteq \frac{K}{k} \left[\left(\frac{\sigma_{bw}^2}{N^2} - \frac{\sigma_{bzw}}{NY} \right) K + \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) \left(\frac{S_{iw}^2}{N^2} - \frac{S_{izw}}{NY} \right) \right] \quad (2)$$

where

$$\sigma_{bzw} = \sum_{i=1}^K (N_i \bar{Z}_i - Y/K) (N_i \bar{W}_i - N/K) / K$$

$$S_{izw} = \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i) (W_{ij} - \bar{W}_i) / (N_i - 1),$$

$$\bar{Z}_i = \sum_{j=1}^{N_i} Z_{ij} / N_i \quad \text{and} \quad \bar{z}_i = \sum_{j=1}^{n_i} z_{ij} / n_i,$$

and σ_{bw}^2 , S_{iw}^2 , \bar{W}_i and \bar{w}_i are the expressions of σ_{bzw} , S_{izw} , \bar{Z}_i and \bar{z}_i respectively, with z replaced by w and Y replaced by N .

Proof. Following a standard result, the relative bias of the estimator \bar{z}_{RS} , to the first order of approximation, is

$$RB(\bar{z}_{RS}) \doteq [V(\hat{N}_{RS})/N^2] - \text{Cov}(\hat{Y}_{RS}, \hat{N}_{RS})/YN. \quad (3)$$

Let E_2 and V_2 denote the conditional expectation and variance for a given sample of clusters and E_1 and V_1 the expectation and variance over all such samples. Then, we have

$$\begin{aligned} V(\hat{N}_{RS}) &= V_1 E_2(\hat{N}_{RS}) + E_1 V_2(\hat{N}_{RS}) \\ &= V_1 \left[\frac{K}{k} \sum_{i=1}^k N_i E_2(\bar{w}_i) \right] \\ &\quad + E_1 \left[\frac{K^2}{k^2} \sum_{i=1}^k N_i^2 V_2(\bar{w}_i) \right] \\ &= V_1 \left(\frac{K}{k} \sum_{i=1}^k N_i \bar{W}_i \right) \\ &\quad + E_1 \left[\frac{K^2}{k^2} \sum_{i=1}^k N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2 \right] \end{aligned}$$

$$= \frac{K^2}{k} \sigma_{bw}^2 + \frac{K}{k} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iw}^2. \quad (4)$$

Similarly, we have

$$\begin{aligned} \text{Cov}(\hat{Y}_{RS}, \hat{N}_{RS}) &= \frac{K^2}{k} \sigma_{bzw} \\ &\quad + \frac{K}{k} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{izw}. \end{aligned} \quad (5)$$

By substituting (4) and (5) in (3), we obtain (2), which completes the proof of the theorem.

Theorem 2. The mean square error (MSE) of the estimator \bar{z}_{RS} , to the first order of approximation, is

$$\text{MSE}(\bar{z}_{RS}) \doteq$$

$$\frac{K}{kN^2} \sum_{i=1}^K N_i^2 \left[(\bar{Z}_i - \bar{Y} \bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right] \quad (6)$$

where $D_i^2 = S_{iz}^2 - 2\bar{Y}S_{izw} + \bar{Y}^2 S_{iw}^2$, and $S_{iz}^2 = \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)^2 / (N_i - 1)$.

Proof. To the first order of approximation, we have

$$\text{MSE}(\bar{z}_{RS}) \doteq [V(\hat{Y}_{RS}) - 2\bar{Y} \text{Cov}(\hat{Y}_{RS}, \hat{N}_{RS}) + \bar{Y}^2 V(\hat{N}_{RS})] / N^2. \quad (7)$$

The expression for $V(\hat{Y}_{RS})$ may be written, following (4), as

$$V(\hat{Y}_{RS}) = \frac{K^2}{k} \sigma_{bz}^2 + \frac{K}{k} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2 \quad (8)$$

where $\sigma_{bz}^2 = \sum_{i=1}^K (N_i \bar{z}_i - Y/K)^2 / K$.

By substituting (4), (5) and (8) in (7), we obtain upon simplification

$$\begin{aligned} \text{MSE}(\bar{z}_{RS}) &\doteq \frac{K^2}{kN^2} (\sigma_{bz}^2 - 2\bar{Y} \sigma_{bzW} + \bar{Y}^2 \sigma_{bW}^2) \\ &+ \frac{K}{kN^2} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) (S_{iz}^2 - 2\bar{Y} S_{izW} + \bar{Y}^2 S_{iW}^2). \quad (9) \end{aligned}$$

Substitution of the expressions for σ_{bz}^2 , σ_{bzW} and σ_{bW}^2 into (9) and simplification yields (6). Now, we provide an estimator of $\text{MSE}(\bar{z}_{RS})$ below.

Theorem 3. A consistent estimator of $\text{MSE}(\bar{z}_{RS})$, to the first order of approximation, is given by

$$\widehat{\text{MSE}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2} \cdot \frac{1}{k-1} \sum_{i=1}^k N_i^2 (\bar{z}_i - \bar{z}_{RS} \bar{w}_i)^2. \quad (10)$$

Proof. We note that the first-stage sampling is done with SRSWR sampling scheme and the random variables $N_i \bar{z}_i$ and $N_i \bar{w}_i$ in the ratio estimator are independently and identically distributed. Hence, the mean square error of \bar{z}_{RS} can be estimated using the well-known result that a variance estimator for a multi-stage design can consider the first stage only (see Särndal, Swensson and Wretman, 1992, Results 2.9.1 and 4.5.1).

From (9), an unbiased estimator of

$$\sigma_{bz}^2 + \frac{1}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2$$

can be written as

$$s_{bz}^2 = \frac{1}{k-1} \sum_{i=1}^k \left(N_i \bar{z}_i - \sum_{i=1}^k N_i \bar{z}_i / k \right)^2, \quad (11)$$

and an unbiased estimator of

$$\sigma_{bzW} + \frac{1}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{izW}$$

is

$$\begin{aligned} s_{bzW} &= \frac{1}{k-1} \sum_{i=1}^k \left(N_i \bar{z}_i - \sum_{i=1}^k N_i \bar{z}_i / k \right) \\ &\times \left(N_i \bar{w}_i - \sum_{i=1}^k N_i \bar{w}_i / k \right). \quad (12) \end{aligned}$$

Similarly, an independent estimator of

$$\sigma_{bW}^2 + \frac{1}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iW}^2$$

is s_{bW}^2 , defined parallel to (11).

Using these results, one can easily show that a consistent estimator of $\text{MSE}(\bar{z}_{RS})$ given in (6) is provided by

$$\widehat{\text{MSE}}(\bar{z}_{RS}) = \frac{K^2}{k\hat{N}_{RS}^2} (s_{bz}^2 - 2\bar{z}_{RS} s_{bzW} + \bar{z}_{RS}^2 s_{bW}^2),$$

which can be written as (10).

2.2 Strategy B

This strategy consists of the following steps:

- Select k clusters out of K by probability proportional to size with replacement (PPSWR) sampling with selection probabilities $P_i = N_i/M$, $i = 1, \dots, K$.
- Same as for strategy A.

Theorem 4. The ratio estimator under PPS sampling

$$\bar{z}_{RP} = \hat{Y}_{RP} / \hat{N}_{RP} = \frac{M}{k} \sum_{i=1}^k \bar{z}_i / \frac{M}{k} \sum_{i=1}^k \bar{w}_i \quad (13)$$

has relative bias, to the first order of approximation,

$$\begin{aligned} RB(\bar{z}_{RP}) &\doteq \frac{M^2}{k} \left[\left(\frac{\sigma_{bW'}^2}{N^2} - \frac{\sigma_{bzW'}}{YN} \right) \right. \\ &\left. + \sum_{i=1}^K \frac{N_i}{M} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) \left(\frac{S_{iW}^2}{N^2} - \frac{S_{izW}}{YN} \right) \right] \quad (14) \end{aligned}$$

where

$$\sigma_{bzW'} = \sum_{i=1}^K (\bar{z}_i - Y/M) (\bar{w}_i - N/M) (N_i/M)$$

and $\sigma_{bW'}^2$ is the expression of $\sigma_{bzW'}$ with z replaced by w and Y replaced by N .

Proof. Using a standard result, the approximate relative bias, to the first order of approximation, is

$$RB(\bar{z}_{RP}) \doteq [V(\hat{N}_{RP})/N^2] - \text{Cov}(\hat{Y}_{RP}, \hat{N}_{RP})/YN. \quad (15)$$

We have

$$\begin{aligned} V(\hat{N}_{RP}) &= V_1 E_2(\hat{N}_{RP}) + E_1 V_2(\hat{N}_{RP}) \\ &= M^2 \left[V_1 \frac{1}{k} \sum_{i=1}^k E_2(\bar{w}_i) + E_1 \frac{1}{k^2} \sum_{i=1}^k V_2(\bar{w}_i) \right] \\ &= M^2 \left[V_1 \frac{1}{k} \sum_{i=1}^k \bar{w}_i + E_1 \frac{1}{k^2} \sum_{i=1}^k \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{izw}^2 \right] \\ &= \frac{M^2}{k} \left[\sigma_{bz'w'}^2 + \sum_{i=1}^K \frac{N_i}{M} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{izw}^2 \right]. \quad (16) \end{aligned}$$

Similarly, one can write

$$\text{Cov}(\hat{Y}_{RP}, \hat{N}_{RP}) = \frac{M^2}{k} \left[\sigma_{bz'w'} + \sum_{i=1}^K \frac{N_i}{M} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{izw} \right]. \quad (17)$$

Substituting (16) and (17) in (15), we obtain (14).

Theorem 5. The MSE of the estimator \bar{z}_{RP} , to the first order of approximation, is

$$\begin{aligned} \text{MSE}(\bar{z}_{RP}) &\doteq \frac{M}{kN^2} \sum_{i=1}^K N_i \\ &\times \left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right]. \quad (18) \end{aligned}$$

Proof. We write, to the first order of approximation,

$$\begin{aligned} \text{MSE}(\bar{z}_{RP}) &\doteq [V(\hat{Y}_{RP}) - 2\bar{Y} \text{Cov}(\hat{Y}_{RP}, \hat{N}_{RP}) \\ &\quad + \bar{Y}^2 V(\hat{N}_{RP})]/N^2. \quad (19) \end{aligned}$$

Also, from Theorem 2.5 of Singh (1988), we have by analogy

$$V(\hat{Y}_{RP}) = \frac{M^2}{k} \sigma_{bz'}^2 + \frac{M^2}{k} \sum_{i=1}^K \frac{N_i}{M} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2, \quad (20)$$

where $\sigma_{bz'}^2 = \sum_{i=1}^K (N_i/M) (\bar{Z}_i - Y/M)^2$. On substituting (16), (17) and (20) in (19) and simplifying, we obtain (18).

Theorem 6. A consistent estimator of $\text{MSE}(\bar{z}_{RP})$, to the first order of approximation, is

$$\widehat{\text{MSE}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}_{RP}^2} \cdot \frac{1}{k(k-1)} \sum_{i=1}^k (\bar{z}_i - \bar{z}_{RP} \bar{w}_i)^2. \quad (21)$$

Proof. As the first-stage units are selected with PPSWR, the justification given in the proof of theorem 3 applies here, as well.

From (20), using Results 2.9.1 and 4.5.1 of Särndal, Swensson and Wretman (1992), an unbiased estimator of

$$\sigma_{bz'}^2 + \sum_{i=1}^K \frac{N_i}{M} \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_{iz}^2$$

can be written as

$$s_{bz'}^2 = \frac{1}{k-1} \sum_{i=1}^k \left(\bar{z}_i - \sum_{i=1}^k \bar{z}_i/k \right)^2. \quad (22)$$

Similarly, defining $s_{bz'w'}$ and $s_{bw'}^2$, one can show that

$$\widehat{\text{MSE}}(\bar{z}_{RP}) = \frac{M^2}{\hat{N}_{RP}^2 k} (s_{bz'}^2 - 2\bar{z}_{RP} s_{bz'w'} + \bar{z}_{RP}^2 s_{bw'}^2),$$

which can be written as (21).

3. EFFICIENCY COMPARISON

The efficiencies of the estimators are compared below under the two strategies.

Remark. The estimator \bar{z}_{RP} under strategy B is expected to be more efficient than the estimator \bar{z}_{RS} under strategy A.

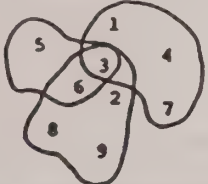
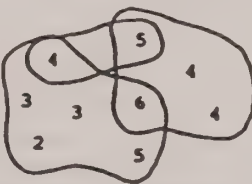
We provide a justification. From (6) and (18), we obtain

$$\begin{aligned} \text{MSE}(\bar{z}_{RS}) - \text{MSE}(\bar{z}_{RP}) &\doteq \frac{M}{kN^2} \sum_{i=1}^K N_i \\ &\times \left[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + \left(\frac{1}{n_i} - \frac{1}{N_i} \right) D_i^2 \right] \left(\frac{KN_i}{M} - 1 \right). \end{aligned}$$

As the cluster size N_i increases, the factor $(KN_i/M - 1)$ will also increase. The other factor of the term under summation is $N_i[(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i)D_i^2]$, which represents the contribution due to variability in z and w present in the i -th cluster (without the constant M/kN^2) towards $\text{MSE}(\bar{z}_{RP})$ in (18). As cluster size N_i increases, the contribution of the i -th cluster towards $\text{MSE}(\bar{z}_{RP})$ is also expected to increase. This makes the covariance between these two factors positive. Hence, the estimator \bar{z}_{RP} is expected to have a smaller MSE than \bar{z}_{RS} .

Table 1

Comparison of the Two Strategies for Two Small Populations

	Population No. 1			Population No. 2		
						
N_i	3	4	5	2	4	6
n_i	1	2	2	1	2	2
Y_{ij}	3,5,6	1,3,4,7	2,3,6,8,9	4,5	4,4,5,6	2,3,3,4,5,6
F_{ij}	3,1,2	1,3,1,1	1,3,2,1,1	2,2	1,1,2,2	1,1,1,2,1,2
Z_{ij}	1,5,3	1,1,4,7	2,1,3,8,9	2,2,5	4,4,2,5,3	2,3,3,2,5,3
W_{ij}	$\frac{1}{3}, 1, \frac{1}{2}$	$1, \frac{1}{3}, 1, 1$	$1, \frac{1}{3}, \frac{1}{2}, 1, 1$	$\frac{1}{2}, \frac{1}{2}$	$1, 1, \frac{1}{2}, \frac{1}{2}$	$1, 1, 1, \frac{1}{2}, 1, \frac{1}{2}$
F	1.38	10.16	18.12	.24	.77	2.94
$MSE(\bar{z}_{RS})$	2.09			0.45		
$MSE(\bar{z}_{RP})$	1.83			0.33		
R.E.	114.21			136.36		
R.B. (\bar{z}_{RS})	-.0105			.0348		
R.B. (\bar{z}_{RP})	-.0047			-.0037		

Numerical Illustration. Here the two proposed sampling strategies are applied to two small populations to shed light on the computations of F_{ij} , Z_{ij} and W_{ij} , and on their comparison. For both the populations $K = 3$, $k = 2$, $M = 12$ and $N = 9$. A unit repeated in two or more clusters represents overlapping. The populations are described in Table 1.

The analysis of the results in Table 1 supports the theoretical developments of the present paper. For both the populations, the factor $F = N_i [(\bar{Z}_i - \bar{Y}\bar{W}_i)^2 + (1/n_i - 1/N_i)D_i^2]$ increases with N_i , resulting in $MSE(\bar{y}_{RP}) < MSE(\bar{y}_{RS})$, as remarked above.

CONCLUSION

This paper removes the realistic limitation of known population size in the earlier work of Singh (1988) while considering overlapping clusters. Also comparison of the two strategies here is more direct, whereas in Singh (1988) the support of evidence given by Hansen and Hurwitz (1943) was needed.

ACKNOWLEDGEMENTS

This research was partially supported by NSERC Grant A-3111. The comments of the referees and the editor on an earlier version were most helpful in improving the paper. These are gratefully acknowledged.

REFERENCES

- AGARWAL, D.K., and SINGH, P. (1982). On cluster sampling strategies using ancillary information. *Sankhyā*, B, 44, 184-192.
- AMDEKAR, S.J. (1985). An unbiased estimator in overlapping clusters. *Bulletin of the Calcutta Statistical Association*, 15, 231-232.
- GIFFARD-JONES, W. (1993). The doctor game. *The Windsor Star*, April 15, 1993.
- GOEL, B.B.P.S., and SINGH, D. (1977). On the formation of clusters. *Journal of the Indian Society of Agricultural Statistics*, 29, 53-68.
- HANSEN, M.H., and HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- SÄRNDAL, C-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, S. (1988). Estimation in overlapping clusters. *Communications in Statistics, Theory and Methods*, 17, 613-621.

PPS Sampling over Two Occasions

N.G.N. PRASAD and J.E. GRAHAM¹

ABSTRACT

The Random Group Method for sampling with probability proportional to size (PPS) is extended to sampling over two occasions. Information on a study variate observed on the first occasion is used to select the matched portion of the sample on the second occasion. Two real data sets are considered for numerical illustration and for comparison with other existing methods.

KEY WORDS: Composite estimator; Efficiency comparisons; Random group method; Probability proportional to size.

1. INTRODUCTION

The practice of using a partial replacement sampling scheme in repeated surveys is quite common due, in part, to an anticipated increase in the efficiency of estimation as well as a reduction in the burden of response. Essentially, after each sampling occasion a fraction of the units observed on that occasion is rotated out of the sample and replaced by a fresh sub-sample from the population. This set of unmatched units is then observed on the next sampling occasion along with the remaining set of matched units. The literature abounds with discussions of sampling and estimation procedures for sampling with equal selection probabilities on two occasions. A particularly important case is the situation where the units are chosen on a given occasion with unequal selection probabilities. In the literature to date, information collected on the previous occasion is used to improve upon the customary estimator of the total or mean for the current occasion by using a difference method of estimation. In this article we present a sampling and estimation procedure for sampling on two occasions which incorporates information collected on the first (previous) occasion in selecting the sub-sample for observation on the second (current) occasion. For the sake of completeness and parsimony, we review only unequal probability selection procedures for two occasions in this section.

Consider a finite population of N units, labelled $1, 2, \dots, N$, and two sampling occasions: 1 (previous occasion) and 2 (current occasion). Let y_{1i} and y_{2i} denote the values of a characteristic y for the i -th unit observed on the first and second occasions respectively. Let Y_1 and Y_2 denote the respective population totals. Suppose a size measure x is known for each of the population units.

1.1 The Des Raj Scheme

Raj (1965) considered the following PPS (probability proportional to size) sampling scheme: On the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values, $i = 1, 2, \dots, N$, and with replacement (wr). On the second occasion a simple random sample s_1 of m units is selected from s without replacement (wor) and an independent PPS sample s_2 of $u = n - m$ units is selected wr from the entire population. Then Y_1 and Y_2 are respectively unbiasedly estimated by:

$$\hat{Y}_1 = \sum_{i \in s} y_{1i} / (np_i) \quad (1.1)$$

and

$$\hat{Y}_2 = Q \hat{Y}_{2u} + (1 - Q) \hat{Y}_{2m}, \quad (1.2)$$

where

$$\hat{Y}_{2u} = \sum_{i \in s_2} y_{2i} / (up_i), \quad (1.3)$$

$$\hat{Y}_{2m} = \sum_{i \in s} y_{1i} / (np_i) + \sum_{i \in s_1} (y_{2i} - y_{1i}) / (mp_i), \quad (1.4)$$

and Q is a weight, $0 \leq Q \leq 1$. Assuming that

$$\begin{aligned} V_1 &= \sum_{i=1}^N (y_{1i}/p_i - Y_1)^2 p_i = V_2 \\ &= \sum_{i=1}^N (y_{2i}/p_i - Y_2)^2 p_i = V, \end{aligned} \quad (1.5)$$

¹ N.G.N. Prasad, Associate Professor, Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta, Canada T6G 2G1; J.E. Graham, Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

the minimum variance of \hat{Y}_2 was found to be

$$V_{\min}(\hat{Y}_2) = V[1 + \sqrt{2(1 - \delta)/(2n)}], \quad (1.6)$$

where δ is given by

$$V\delta = \sum_{i=1}^N (y_{1i}/p_i - Y_1)(y_{2i}/p_i - Y_2)p_i. \quad (1.7)$$

1.2 The Ghangurde-Rao (G-R) Scheme

Under the PPSWOR framework, Ghangurde and Rao (1969) extended the Rao-Hartley-Cochran (RHC) Method, also known as the Random Group Method (See: Rao, Hartley and Cochran 1962) to sampling on two occasions. Under the RHC Method, the population of N units is split at random into n groups of sizes N_1, N_2, \dots, N_n such that $\sum_{h=1}^n N_h = N$, and a sample of one unit is drawn independently from each of the n groups with probabilities proportional to the initial selection probabilities, p_i . Under the G-R Method, the population is first divided at random into n groups, each of size N/n (assumed to be an integer). On the first occasion, one unit is drawn from each random group (as described above), giving a sample s of n units. On the second occasion, a simple random wor sample s_1 of $m = \lambda n$ ($0 < \lambda < 1$) matched units is selected from s and an independent sample s_2 of $u = n - m$ units is drawn from the whole population of N units by the same method that was used in obtaining s . Then, a composite estimator of Y_2 is given by

$$\hat{Y}'_2 = Q' \hat{Y}'_{2u} + (1 - Q') \hat{Y}'_{2m}, \quad (1.8)$$

where $0 \leq Q' \leq 1$,

$$\hat{Y}'_{2u} = \sum_{i \in s_2} \frac{y_{2i} P_i^*}{p_i}, \quad (1.9)$$

and

$$\hat{Y}'_{2m} = \sum_{i \in s} \frac{y_{1i} P_i}{p_i} + nm^{-1} \sum_{i \in s_1} \frac{(y_{2i} - y_{1i}) P_i}{p_i}, \quad (1.10)$$

with P_i and P_i^* denoting the totals of the p_i values for the groups containing the i -th unit ($i = 1, 2, \dots, N$) in the selection of s and s_2 respectively. Under assumption (1.5), the variance of \hat{Y}'_2 (with optimum values of Q' and λ) is given by

$$V_{\min}(\hat{Y}'_2) = \frac{NV}{2n(N-1)} \times [1 - n/N + \sqrt{2(1 - \delta)}(1 + \gamma)n/N], \quad (1.11)$$

where

$$\gamma = \frac{(1 - \rho')V'}{(1 - \delta)V} - 1,$$

$$V' = N^{-1} \sum_{i=1}^N (y_{1i} - \bar{Y}_1)^2 = N^{-1} \sum_{i=1}^N (y_{2i} - \bar{Y}_2)^2$$

and

$$\rho' = N^{-1} \sum_{i=1}^N (y_{1i} - \bar{Y}_1)(y_{2i} - \bar{Y}_2)/V'.$$

1.3 The Chotai Scheme

Chotai (1974), under the additional assumption that n/m is an integer, modified the G-R sampling design on the second occasion. A sample s is selected as in the G-R scheme on the first occasion. On the second occasion, the n units in the sample s are split at random into $m (= \lambda n)$ groups each of size n/m . One unit is selected from each of the m groups independently with probabilities proportional to the P_i 's as defined in the G-R scheme. This selection yields the sample s_1 . The selection of s_2 is the same as in the G-R scheme. Then a composite estimator of Y_2 is given by

$$\hat{Y}_2^C = Q^C \hat{Y}_{2u}^C + (1 - Q^C) \hat{Y}_{2m}^C, \quad (1.12)$$

where $0 \leq Q^C \leq 1$,

$$\hat{Y}_{2u}^C = \sum_{i \in s_2} \frac{y_{2i} P_i^*}{p_i}, \quad (1.13)$$

and

$$\hat{Y}_{2m}^C = \sum_{i \in s_1} \frac{(y_{2i} - y_{1i}) P_i^+}{p_i} + \sum_{i \in s} \frac{y_{1i} P_i}{p_i}. \quad (1.14)$$

Here, P_i and P_i^* are as defined in the G-R scheme, and P_i^+ denotes the total of the P_i -values for those random groups of s containing the i -th unit ($i = 1, 2, \dots, N$) in the selection of s_1 . The minimum variance of \hat{Y}_2^C under assumption (1.5), obtained by using the optimum values of Q^C and λ , is given by

$$V_{\min}(\hat{Y}_2^C) = \frac{NV}{2n(N-1)} [1 - n/N + \sqrt{2(1 - \delta)}]. \quad (1.15)$$

Under this scheme, but without assumption (1.5), Chotai also considered an estimator of Y_2 (similar to Kulldorff's estimator for simple random sampling: See Kulldorff 1963), given by

$$\hat{Y}_2^{CM} = Q^{CM} \hat{Y}_{2u}^C + (1 - Q^{CM}) \hat{Y}_{2m}^{CM}, \quad (1.16)$$

where \hat{Y}_{2u}^C is as defined in (1.13), Q^{CM} ($0 \leq Q^{CM} \leq 1$) is an assigned weight to be determined and

$$\hat{Y}_{2m}^{CM} = \sum_{i \in s_1} \frac{(y_{2i} - \beta y_{1i}) P_i^+}{p_i} + \beta \sum_{i \in s} \frac{y_{1i} P_i}{p_i}, \quad (1.17)$$

with

$$\beta = \delta \left[\frac{\sum_{i=1}^N p_i (y_{2i}/p_i - Y_2)^2}{\sum_{i=1}^N p_i (y_{1i}/p_i - Y_1)^2} \right] = \delta \frac{V_2}{V_1}, \quad (1.18)$$

and δ as defined in (1.7). The minimum variance of \hat{Y}_2^{CM} , using optimum values of Q^{CM} and λ , is given by

$$V_{min}(\hat{Y}_2^{CM}) = \frac{N}{2n(N-1)} (1 + \sqrt{1 - \delta^2} - n/N) V_2. \quad (1.19)$$

To actually use \hat{Y}_2^{CM} it is evidently necessary to first assess the value of β , which is usually not possible in practice. An estimate of β based on the available sample can be used but this will induce a bias in the estimator \hat{Y}_2^{CM} .

2. ALTERNATIVE SCHEMES FOR SAMPLING PPS OVER TWO OCCASIONS

We now present an alternative sampling and estimation procedure which does not require a known value of β as defined in (1.18). In this scheme information collected on the first occasion is used in selecting the sample on the second occasion. The approach is based upon a procedure developed by Prasad and Srivenkataramana (1980) and was used there in the context of double sampling where a second phase sub-sample is selected using information obtained from an initial sample. For simplicity, we first consider its implementation in Raj's (1965) scheme (described earlier).

2.1 A Modification of Des Raj's Scheme

On the first occasion a sample s of size n is selected with probabilities p_i proportional to the x_i values and with replacement. On the second occasion, instead of choosing

a SRSWOR sub-sample, a sub-sample s_1 of m units is selected from s using a PPSWR scheme with size measure $z_i = y_{1i}/x_i$, where y_{1i} is the observed value for the y characteristic for unit i on the first occasion. A sample s_2 of size $u = n - m$ is drawn, independent of s , as in Raj (1965). A composite estimator of Y_2 is given by

$$\tilde{Y}_2 = Q \hat{Y}_{2u} + (1 - Q) \tilde{Y}_{2m},$$

where \hat{Y}_{2u} is as defined in (1.3) and

$$\tilde{Y}_{2m} = \frac{1}{nm} \sum_{i \in s_1} \frac{(y_{2i}/p_i)}{(y_{1i}/p_i)} \sum_{i \in s} (y_{1i}/p_i),$$

with Q being a weight, $0 \leq Q \leq 1$. The minimum variance of \tilde{Y}_2 , obtained by minimizing the variance of \tilde{Y}_2 with respect to Q , is given by

$$V_{min}(\tilde{Y}_2) = V_1 C_1 (n + C_1 m)^{-1},$$

where $C_1 = \sum_{i=1}^N (y_{2i}/p_{1i} - Y_2)^2 p_{1i} V_1^{-1}$, with $p_{1i} = y_{1i}/Y_1$ and V_1 as defined in (1.5).

2.2 A Modification to Chotai's Scheme

As in Chotai (1974), assume that N , n , and $m (< n)$ are all positive integers such that N/n , N/u and n/m are also all integers. Then:

1. For the first occasion select a sample s of size n in the same manner as that adopted in the G-R procedure. For this set of units, observations y_{1i} , $i = 1, \dots, n$, are made on a characteristic y .
2. For the second occasion, (a) split the n units in s at random into m groups, each of size n/m and draw one unit with PPS, $p_i^* = (y_{1i} P_i)/p_i$, independently from each of the m groups, yielding a sub-sample s_1 , where P_i is as defined in the G-R scheme; (b) select s_2 , a fresh sample of $u = n - m$ units from the entire population, and observe the second occasion y values, y_{2i} , for these u units in the same manner as in the G-R scheme.

Note that the difference between the proposed procedure and that of Chotai (1974) lies in the selection of s_1 : in the former, information collected on the first occasion is used in selecting s_1 on the second occasion.

We now consider an estimator of the second occasion total Y_2 that exploits the proposed procedure. Let

$$y_{2i}^* = \frac{y_{2i} P_i}{p_i}.$$

A composite estimator of Y_2 is given by

$$\hat{Y}_2^* = Q^{**} \hat{Y}_{2u}^C + (1 - Q^{**}) \hat{Y}_{2m}^*, \quad (2.1)$$

where \hat{Y}_{2u}^C is defined as in (1.13), $0 \leq Q^{**} \leq 1$ and

$$\hat{Y}_{2m}^* = \sum_{i \in s_1} \frac{y_{2i}^* \tilde{P}_i}{p_i^*}.$$

Here \tilde{P}_i denotes the total of the p_i^* values associated with those units that belong to the random group from which the i -th unit was selected in s_1 . Let E_1 and E_2 denote expectation and V_1 and V_2 denote variance over all s and for a given s , respectively. The unbiasedness of \hat{Y}_{2m}^* and hence of \hat{Y}_2^* for Y_2 follows by noting that the expected value of \hat{Y}_{2m}^* is

$$E(\hat{Y}_{2m}^*) = E_1 E_2(\hat{Y}_{2m}^*) = E_1 \left(\sum_{i \in s} \frac{y_{2i} P_i}{p_i} \right) = Y_2. \quad (2.2)$$

To obtain the variance of \hat{Y}_{2m}^* , consider

$$\begin{aligned} V_2(\hat{Y}_{2m}^*) &= \frac{n-m}{m(n-1)} \sum_{i \in s} \left(\frac{y_{2i}^*}{p_i^*} - \sum_{i \in s} y_{2i}^* \right)^2 p_i^* \\ &= \frac{n-m}{m(n-1)} \left[\sum_{i \in s} \frac{(y_{2i}^2/y_{1i})}{p_i} P_i \sum_{i \in s} \frac{y_{1i} P_i}{p_i} \right. \\ &\quad \left. - \left(\sum_{i \in s} \frac{y_{2i} P_i}{p_i} \right)^2 \right], \end{aligned}$$

which leads, after considerable algebraic simplification, to

$$E_1 V_2(\hat{Y}_{2m}^*) = \frac{N(n-m)}{mn(N-1)} \sigma_2^2,$$

where

$$\sigma_2^2 = \sum_{i=1}^N \left(\frac{y_{2i}}{y_{1i}} Y_1 - Y_2 \right)^2 \frac{y_{1i}}{Y_1}.$$

Noting that

$$V_1 E_2(\hat{Y}_{2m}^*) = \frac{N-n}{n(N-1)} \sigma_2^2,$$

it follows that

$$V(\hat{Y}_{2m}^*) = \frac{N}{n(N-1)} \left[(1 - n/N) + \frac{1-\lambda}{\lambda} h \right] \sigma_2^2, \quad (2.3)$$

where

$$h = \frac{\sigma_3^2}{\sigma_2^2}, \quad \sigma_2^2 = V_2 = \sum_{i=1}^N (y_{2i}/p_i - Y_2)^2 p_i \quad \text{and} \quad \lambda = \frac{m}{n}.$$

Because \hat{Y}_{2u}^C and \hat{Y}_{2m}^* are independent, the variance of \hat{Y}_2^* is given by

$$V(\hat{Y}_2^*) = Q^{**2} V(\hat{Y}_{2u}^C) + (1 - Q^{**})^2 V(\hat{Y}_{2m}^*),$$

where

$$V(\hat{Y}_{2u}^C) = \frac{N-u}{u(N-1)} \sigma_2^2,$$

and $V(\hat{Y}_{2m}^*)$ is given by (2.3).

The minimum variance of $V(\hat{Y}_2^*)$ is obtained by using optimum values of Q^{**} and λ , respectively given by

$$Q^{**} = \frac{(1 - n/N) + \frac{(1-\lambda)}{\lambda} h}{(1 - n/N) + \frac{(1-\lambda)}{\lambda} h + \frac{(1 - (1-\lambda)n/N)}{(1-\lambda)}},$$

and

$$\lambda = \frac{\sqrt{h}}{1 + \sqrt{h}}.$$

Hence, the minimum variance of $V(\hat{Y}_2^*)$ is given by

$$V_{min}(\hat{Y}_2^*) = \frac{N\sigma_2^2}{n(N-1)} [1 - n/N + \sqrt{h}]. \quad (2.4)$$

Note that the quantity h reflects the efficiency of the estimator using the p_i 's as initial selection probabilities over the estimator with initial selection probabilities y_{1i}/Y_1 . A "small" value of h leads to an increase in the efficiency of the proposed method over Chotai's.

3. NUMERICAL EFFICIENCY COMPARISONS

The composite estimators \hat{Y}_2^C defined in (1.12), \hat{Y}_2^{CM} defined in (1.16) and \hat{Y}_2^* defined in (2.1) are now compared at their respective optimum Q and λ values. The efficiency of the scheme proposed in 2.2 relative to Chotai's (1974) procedure is examined through a comparison of the following two relative efficiencies:

$$RE1 = \frac{V_{min}(\hat{Y}_2^C)}{V_{min}(\hat{Y}_2^*)} = \frac{(1 - n/N) + \sqrt{2(1-\delta)}}{(1 - n/N) + \sqrt{h}}$$

and

$$RE2 = \frac{V_{min}(\hat{Y}_2^{CM})}{V_{min}(\hat{Y}_2^*)} = \frac{(1 - n/N) + \sqrt{1-\delta^2}}{(1 - n/N) + \sqrt{h}},$$

evaluated respectively obtained using (1.15) and (2.4), and (1.19) and (2.4). It follows that the proposed scheme is superior to that of Chotai using Kulldorff's estimator (which depends on the unknown constant β) for those populations having $h < (1 - \delta^2)$. In order to permit meaningful numerical comparisons, two data sets that have appeared elsewhere in the literature are used here.

Data Set A: This data set relates to the area under wheat in 1964 (y_2), in 1963 (y_1) and cultivated area in 1961 (x) for 34 villages in India (See Murthy 1967). The parameter values for this data set are $\delta = 0.6404$ and $h = 0.1868$.

Data Set B: This data set relates to the area under wheat in 1937 (y_2) and in 1936 (y_1) and cultivated area in 1930 (x) for a sample of 34 villages in India (see: Sukhatme, P.V. and Sukhatme, B.V. 1970). The corresponding parameter values for this data set are $\delta = 0.7635$ and $h = 0.3811$.

Using these values for δ and h the two relative efficiencies values RE1 and RE2 (expressed as percentages) were computed for selected values of n/N and are given in Tables 1 and 2.

Table 1
RE1% - Values for Data Sets A and B

n/N	Data Set A	Data Set B
0.05	130.09	124.30
0.10	131.22	125.21
0.15	132.43	126.19
0.20	133.75	127.25
0.25	135.18	128.41
0.30	136.73	129.66

Table 2
RE2% - Values for Data Sets A and B

n/N	Data Set A	Data Set B
0.05	104.49	101.82
0.10	104.64	101.88
0.15	104.80	101.94
0.20	104.97	102.01
0.25	105.15	102.08
0.30	105.34	102.16

An examination of Table 1 leads to the conclusion that the proposed scheme out performs that of Chotai (1974). The gain in the efficiency ranges from 30% to 37% for Data Set A and from 24% to 30% for Data Set B as the sampling fraction varies from 0.05 to 0.30. Note that the increase in efficiency is greater for Data Set A than for Data Set B because of the difference in the value of the

parameters h (0.1868 vs. 0.3811) and of δ (0.6404 vs. 0.7635). Recall that h measures the efficiency of p_i as a size measure for unit i compared to the use of y_{1i} as a size measure in estimating the total Y_2 for the current occasion and δ is the correlation between y_{1i}/p_i and y_{2i}/p_i as defined in (1.7). When h is relatively small, greater gains in efficiency are realized with the proposed scheme than when h is not small. In both cases, however, the efficiency gains using the proposed procedure are worthwhile.

The efficiency gains using the proposed method compared to the use of Chotai's scheme with Kulldorff's estimator (as reported in Table 2) are minimal, varying from 4.5% to 5.3% for Data Set A and from 1.8% to 2.2% from Data Set B. But in order to use Kulldorff's estimator, the value of β must be available. In practice this is not the case. It follows that the proposed strategy performs well from the point of view of actual implementation and of efficiency gain.

There are situations where the auxiliary information needed to compute the initial selection probabilities is not available. A simple random sampling scheme may then be used in place of the RHC procedure in selecting the sample for the first occasion enumeration; the RHC procedure can then be adopted in selecting s_1 by using the SRS information on the study variable collected on the first occasion. The theory for such a procedure follows directly as a special case of that presented by taking $p_i = 1/N, i = 1, \dots, N$. One would anticipate that substantial gains in efficiency would then result in this situation.

ACKNOWLEDGEMENTS

The authors thank Professor J.N.K. Rao for suggesting this problem, and the referee for constructive suggestions. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Series C, 36, 173-180.

GHANGURDE, P., and RAO, J.N.K. (1969). Some results on sampling over two occasions. *Sankhyā*, Series A, 31, 463-472.

KULLDORFF, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Review of the International Statistical Institute*, 31, 24-57.

MURTHY, N.N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.

PRASAD, N.G.N., and SRIVENKATARAMANA, T. (1980). Double sampling with PPS selection. *Vignana Bharathi*, 6, 52-58.

RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.

SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys With Applications*. Ames, Iowa: Iowa State University Press.

Multi-way Stratification by Linear Programming

R.R. SITTER and C.J. SKINNER¹

ABSTRACT

Rao and Nigam (1990, 1992) showed how a class of controlled sampling designs can be implemented using linear programming. In this article their approach is applied to multi-way stratification. A comparison is made with existing methods both by illustrating the sampling schemes generated for specific examples and by evaluating mean squared errors. The proposed approach is relatively simple to use and appears to have reasonable mean squared error properties. The computations required can, however, increase rapidly as the number of cells in the multi-way classification increase. Variance estimation is also considered.

KEY WORDS: Controlled selection; Linear programming; Multistage sampling; Stratified sampling.

1. INTRODUCTION

There are often several stratifying variables available to the sample designer and it is natural in such cases for the designer to consider defining strata as the cells formed by cross-classifying categories of these variables. A problem with this approach, particularly common when selecting primary sampling units (psu's) in household surveys, is that the desired sample size may be less than the total number of cells and hence conventional methods of stratification may be inapplicable.

An illustration, based on a hypothetical example of Bryant *et al.* (1960), is given in Table 1. Communities (psu's) are classified by two stratifying factors: type of community with three categories and region with five categories. The desired sample size of $n = 10$ is less than the total number of cells, 15. This example also illustrates a related problem. The entries in Table 1 are the expected counts under proportionate stratification, that is the population proportions multiplied by the sample size. Even if the sample size was doubled to exceed the number of cells, the expected sample counts would still not be integers. Whilst the effect of rounding such values to integers may not be practically significant for large expected counts, the choice of how to round with very small expected counts may be of greater concern.

One reaction to the problem of many cells is simply to drop one or more of the stratifying variables or to group some of the categories. Alternatively, a number of procedures have been proposed which attempt to retain some 'control' for all the categories of all the stratifying variables by permitting different forms of random selection of cells.

Goodman and Kish (1950) proposed one procedure under the title 'controlled selection'. Jessen (1970) suggests that 'this method is somewhat complicated and its use in applied sampling appears limited' (p. 778). Waterton (1983)

Table 1

Expected Sample Cell Counts Under Proportionate Stratification with $n = 10$

Regions	Type of Community			Total
	Urban	Rural	Metropolitan	
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

illustrates this complexity. Bryant *et al.* (1960) propose a much simpler method for two-way stratification. Their method has the property that the expected sample counts display independence between the rows and columns of the two-way table. If the rows and columns are also independent in the population then there is no problem but if, as will often be the case, there is an appreciable lack of independence then some reweighting will usually be necessary and this can be unattractive in practice and can inflate the variance as is shown in Section 5. Jessen (1970) points out that a further limitation of the method of Bryant *et al.* (1960) is that it is not possible to constrain specified cell sizes to be zero. He proposes two approaches for both two-way and three-way stratification but both approaches remain fairly complicated to implement and, as noted by Causey *et al.* (1985), do not always lead to a solution.

All the above methods may be carried out by hand with varying degrees of laboriousness, but none take advantage of the power of modern computing. In this paper we shall show how computational procedures of linear programming can be applied to the multi-way stratification problem following Rao and Nigam (1990, 1992). Our proposed approach may be viewed as complementing the linear

¹ R.R. Sitter, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6; C.J. Skinner, Department of Social Statistics, University of Southampton SO9 5NH, U.K.

programming approach proposed by Causey *et al.* (1985). Which of the two approaches is preferable will depend on the nature of the stratification problem and on the software available. The potential disadvantage of our approach is that it can be much more computationally intensive, since the number of unknowns in our linear programming problem may be as large as $\binom{k}{n}$, when k is the number of cells in the table and n is the sample size, whereas the number of unknowns in the approach of Causey *et al.* (1985) is only k . A number of suggestions will be made, however, to reduce the computational demands of our approach. There are several potential advantages of our approach. First, the stratification problem corresponds directly to the linear programming problem and so the computer programming is straightforward, whereas the approach of Causey *et al.* is less direct, involving mimicking the behaviour of nonlinear functions by linear functions (p. 904) and nesting repeated linear programming problems within a further recursive algorithm. Second, our procedure always has a solution, whereas the procedure of Causey *et al.* need not, for example in cases of three-way stratification. Third, the objective function in our linear programming problem can be naturally modified to reflect the different objectives of the stratification problem, for example in a three-way problem where it is more important to 'balance' the sample with respect to the first two stratifying variables than the third. Fourth, our procedure can be naturally modified to constrain the joint inclusion probabilities of cells to be positive in order to permit unbiased variance estimation.

2. THE PROPOSED APPROACH

2.1 Basic Ideas

We begin with the simplest kind of two-way stratification. Let a population of N units be classified into the RC cells of a two-way table formed by cross-classifying a row stratification factor with R categories and a column factor with C categories. Let N_{ij} be the number of units in cell ij , that is the set of units in both row i and column j , and let $P_{ij} = N_{ij}/N$ be the corresponding proportion. The parameter of interest is taken to be the population mean, \bar{Y} , of a variable Y .

Consider the following two-stage sampling procedure. First, sample sizes n_{ij} are determined for each cell according to a specified randomized procedure. Letting s denote the $R \times C$ array (n_{ij} , $i = 1, \dots, R, j = 1, \dots, C$), this procedure assigns a probability $p(s)$ to each s in a set S of possible arrays. To emphasize the dependence of n_{ij} on s we write $n_{ij}(s)$. Second, a simple random sample of $n_{ij}(s)$ units is selected from cell ij and the values of Y are recorded for the sample units.

We restrict attention to designs of fixed sample size $n > 0$, that is we restrict S to be the set S_n of all arrays such that

$$\sum_{i=1}^R \sum_{j=1}^C n_{ij}(s) = n.$$

We also restrict attention to proportionate stratification so that

$$\sum_{s \in S_n} n_{ij}(s)p(s) = nP_{ij} \quad \text{for } i = 1, \dots, R, \\ j = 1, \dots, C. \quad (2.1)$$

It follows from (2.1) that the simple unweighted sample mean $\bar{y}(s)$ is an unbiased estimator of \bar{Y} . We propose to choose a (or the) sampling design $p(s)$ which minimizes the expected lack of 'desirability' of the sample s by solving the problem:

$$\text{minimize } \sum_{s \in S_n} w(s)p(s), \quad (2.2)$$

subject to the constraint (2.1), where $w(s)$ is a loss function for the sample s to be specified and P is the class of possible sample designs on S_n obeying

$$0 \leq p(s) \leq 1 \quad \text{for all } s \in S_n. \quad (2.3)$$

Note that (2.1) implies $\sum_{s \in S_n} p(s) = 1$. The key observation of Rao and Nigam (1990, 1992) is that the objective function in (2.2) and the equality and inequality constraints in (2.1) and (2.3) are all linear in $p(s)$ and hence this problem may be solved directly by linear programming with the $p(s)$, $s \in S_n$, as unknowns. The main obstacle to this approach is that the number of elements in S_n is often very large and even with modern computing power it becomes difficult to carry out linear programming if the number of unknowns is large.

It is therefore desirable to restrict attention to a subset of S_n . One natural restriction is to consider only arrays s for which $n_{ij}(s)$ is either equal to $I_{ij} = [nP_{ij}]$, the greatest integer less than nP_{ij} , or $I_{ij} + 1$. Letting $\tilde{n}_{ij}(s) = n_{ij}(s) - I_{ij}$ and $r_{ij} = nP_{ij} - I_{ij}$ the problem becomes

$$\text{minimize } \sum_{s \in \tilde{S}_n} w(s)p(s), \quad (2.4)$$

subject to

$$\sum_{s \in \tilde{S}_n} \tilde{n}_{ij}(s)p(s) = r_{ij}, \quad (2.5)$$

$$\sum_{s \in \tilde{S}_n} p(s) = 1, \quad 0 \leq p(s) \leq 1 \quad \text{for all } s \in \tilde{S}_n, \quad (2.6)$$

where $\tilde{S}_{\tilde{n}}$ is the set of $R \times C$ arrays, where all elements are 0 or 1 and the sum of elements is $\tilde{n} = n - \sum_{ij} I_{ij}$. Note, of course, that if all the I_{ij} are zero, then this is just the same problem as before. The number of elements in $\tilde{S}_{\tilde{n}}$, which determines the magnitude of the computational task for linear programming, is now $\binom{RC}{\tilde{n}}$. This number can still be very large, however, and some further reduction can be achieved by sensible choice of the loss function $w(s)$ as discussed in the next section.

For Table 1, this would amount to considering the situation represented by Table 2, while only allowing a 0 or 1 cell sample size, and then adding back 1 to cells (1,1), (3,3), (4,2) and (5,1) in the final solution. Thus $n = 10$, but $\tilde{n} = 6$.

Table 2

Table of r_{ij} Values from Table 1 with $\tilde{n} = 6$

Regions	Type of Community			Total
	Urban	Rural	Metropolitan	
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

2.2 Choice of Loss Function $w(s)$

The major flexibility of the proposed approach derives from the user's freedom to choose the function $w(s)$ which enters the objective function in (2.2). The conventional approach to two-way stratification (e.g., Jessen 1970; Causey *et al.* 1985) is to require that the selected sample s obey the marginal constraints:

$$|n_{i\cdot}(s) - nP_{i\cdot}| < 1 \quad i = 1, \dots, R, \quad (2.7)$$

$$|n_{\cdot j}(s) - nP_{\cdot j}| < 1 \quad j = 1, \dots, C, \quad (2.8)$$

where

$$n_{i\cdot}(s) = \sum_j n_{ij}(s), \quad n_{\cdot j}(s) = \sum_i n_{ij}(s)$$

$$P_{i\cdot} = \sum_j P_{ij}, \quad P_{\cdot j} = \sum_i P_{ij}.$$

This requirement can be accommodated in our approach by setting $w(s)$ as (effectively) infinite for samples s not satisfying (2.7) or (2.8) or more simply by excluding such samples from the set S_n . The problem with this conventional approach is that no solution to the constrained optimization-problem may exist.

In our approach, however, if we use a loss function such as

$$w(s) = \sum_{i=1}^R (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^C (n_{\cdot j}(s) - nP_{\cdot j})^2, \quad (2.9)$$

then an optimal solution will always exist within a large enough set S_n . In practice, it may be advantageous computationally to restrict the set S_n initially to only those samples obeying (2.7) and (2.8), or even a subset of these, and then to expand the set if necessary, say by changing 1 to 2 in (2.7) and (2.8), until a solution is found.

Let us now consider the more fundamental question of why constraints such as (2.7) and (2.8) are sensible anyway. From a non-statistical point of view, the balancing of a sample with respect to factors with a known population distribution may reassure users about the 'representativeness' of the sample. From a statistical point of view, given our unbiasedness constraint (2.1), it is natural to consider how the loss function might be chosen to improve efficiency. This question may be examined by taking $w(s)$ as the mean squared error $E_m(\bar{y}(s) - \bar{Y})^2$ under a model m . Then the solution to the optimization problem (2.2) minimizes the design-expected model-mean squared error or equivalently, since we require design-unbiasedness, the model-expected design variance.

Consider, for example, the main-effects analysis of variance model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

where y_{ijk} is the k th value of Y in cell ij , μ is a fixed mean and α_i , β_j and ϵ_{ijk} are independent zero-mean random effects with variances σ_α^2 , σ_β^2 and σ_ϵ^2 , respectively. Then, ignoring finite population correction terms,

$$E_m(\bar{y}(s) - \bar{Y})^2 = \sigma_\alpha^2 \sum_i (n_{i\cdot}(s)/n - P_{i\cdot})^2 + \sigma_\beta^2 \sum_j (n_{\cdot j}(s)/n - P_{\cdot j})^2 + \sigma_\epsilon^2/n. \quad (2.10)$$

Hence, if $\sigma_\alpha^2 = \sigma_\beta^2$ the expected design variance of $\bar{y}(s)$ under this model is minimized by taking the loss function in (2.9). Alternatively, if one had some prior information about the likely ratio of the between row variance relative to the between column variance then it may be sensible, on efficiency grounds, to modify the loss function in (2.9) by multiplying the first term on the right hand side of (2.9) by this estimated ratio.

On the other hand if it is thought *a priori* that there is likely to be a strong interaction between the row and column factors in their effect on Y then simply attempting to balance on the margins may be inappropriate. For

example, if one stratification factor is urban/rural and the other is an economic indicator X and it is known that Y is positively related to X in urban areas and negatively related in rural areas then it is likely to be more efficient to stratify partially by X separately within rural and urban areas than to balance fully on both margins. See Bryant *et al.* (1960, section 9) for related comments on efficiency for two-way stratification.

2.3 Higher-way Stratification

The proposed approach extends naturally to 3 or more stratifying factors by letting s denote the corresponding r -way array. The loss function will typically include further terms, for example for three-way stratification we might take

$$w(s) = \lambda_1 \sum_{i=1}^{R_1} (n_{i..}(s) - nP_{i..})^2 + \lambda_2 \sum_{j=1}^{R_2} (n_{.j.}(s) - nP_{.j.})^2 + \lambda_3 \sum_{k=1}^{R_3} (n_{..k}(s) - nP_{..k})^2$$

in obvious notation, where λ_1 , λ_2 and λ_3 are included to represent the relative importance of balancing on the three factors and might consist of prior estimates of the variances of the Y means between categories of the three stratifying factors, as in (2.10).

2.4 Multistage Sampling

One important practical application of multi-way stratification is to the selection of primary sampling units (psu's) in multistage sampling, where it is common for information of several stratifying factors to be available.

In the approach of Section 2.1, the inclusion probabilities of each population unit are $E(n_{ij}(s)/N_{ij}) = n/N$. If it is desired to select psu's with equal probability then this approach extends immediately with the psu's constituting the units and with the observed values of Y replaced by unbiased estimators of the psu totals. Suppose instead that it is desired to select psu's with unequal probabilities, say nz_{ijk} for psu k in cell ij , where usually z_{ijk} will equal $M_{ijk}/\sum_{jk} M_{ijk}$, with M_{ijk} being some measure of size of psu k in cell ij . Then the procedure may be simply modified by setting P_{ij} equal to the sum of z_{ijk} over psu's k in cell ij . Then, if $n_{ij}(s) > 0$, a sample of $n_{ij}(s)$ psu's in cell ij is selected by some probability proportional to z_{ijk} method.

3. EXAMPLES

Example 1: Bryant, Hartley and Jessen (1960)

We will first demonstrate the method on the hypothetical example of Bryant *et al.* (1960) given in Table 1. We first reduce the problem to the form of (2.4), (2.5) and (2.6), where the r_{ij} 's are given in Table 2. The weight function in (2.9) in this reduced linear programming problem becomes

$$w(s) = \sum_{i=1}^5 (\tilde{n}_{i.}(\tilde{s}) - r_{i.})^2 + \sum_{j=1}^3 (\tilde{n}_{.j}(\tilde{s}) - r_{.j})^2.$$

Applying a standard linear programming package in the NAG FORTRAN library, we obtain the solution given in Table 3. The I_{ij} values have been added to the solution so that $n_{ij} = I_{ij} + \tilde{n}_{ij}(\tilde{s})$. It turns out for this solution that each s , for which $p(s) > 0$, has margins $n_{i.}(s)$ and $n_{.j}(s)$ which match the desired margins exactly, that is the solution makes (2.4) zero.

Table 3
Solution to Example 1

s	$p(s)$	s	$p(s)$
1 1 0		1 1 0	
1 0 0		0 0 1	
0 1 1	0.2	0 1 1	0.1
0 2 1		1 1 1	
1 0 1		1 1 0	
1 1 0		1 0 1	
0 0 1		0 1 0	
0 0 2	0.2	1 0 1	0.2
1 2 0		0 2 1	
1 1 0		1 1 0	
1 0 1		1 0 1	
0 1 0		0 0 1	
0 1 1	0.1	0 1 1	0.2
1 1 1		1 2 0	
1 1 0		1 1 0	

Example 2: Jessen (1970)

Jessen (1970) proposed two methods for two-way and three-way stratification. Both of these are quite complicated and involve determining the set of samples which exactly match the margins. Neither method is guaranteed to yield a solution. Jessen (1970) applies both methods to a simple hypothetical example for which both yield a solution. This example is reproduced in Table 4. In this example, since all of the $nP_{ij} < 1$, the linear programming problems defined by (2.1), (2.2) and (2.3) and by (2.4), (2.5) and (2.6), respectively, are identical. We applied our method to this problem, again using the $w(s)$ as defined in (2.9). By trying a number of different seeds

in the optimization routine, we were able to obtain three different solutions, all of which make (2.2) zero and satisfy the constraints. These are given in Table 5. The first two solutions are the same two as obtained by Jessen's method 2 and method 3, respectively.

Table 4
Example 2: Jessen (1970)
Expected Sample Cell Counts Under Proportionate Stratification with $n = 6$

Rows	Columns			$nP_{i\cdot}$
	1	2	3	
1	0.8	0.5	0.7	2.0
2	0.7	0.8	0.5	2.0
3	0.5	0.7	0.8	2.0
$nP_{\cdot j}$	2.0	2.0	2.0	6.0

Table 5
Solution to Example 2

s	$p_1(s)$	$p_2(s)$	$p_3(s)$
1 0 1	0.5	0.4	0.3
1 1 0			
0 1 1			
1 1 0	0.3	0.2	0.1
0 1 1			
1 0 1			
0 1 1	0.2	0.1	0.0
1 0 1			
1 1 0			
1 1 0	0.0	0.1	0.2
1 0 1			
0 1 1			
1 0 1	0.0	0.1	0.2
0 1 1			
1 1 0			
0 1 1	0.0	0.1	0.2
1 1 0			
1 0 1			

Example 3: Causey, Cox and Ernst (1985)

Causey *et al.* (1985) give an example of three-way stratification for which their method fails to yield a solution. They consider a population subject to a $2 \times 2 \times 2$ stratification from which a sample of size $n = 2$ is to be drawn, with the expected sample size in the ijk -th cell, n_{ijk} , as follows:

$$n_{111} = n_{221} = n_{122} = n_{212} = .5$$

$$n_{121} = n_{211} = n_{112} = n_{222} = 0.$$

If we apply our method in a similar manner to Examples 1 and 2 we obtain the solution given in Table 6. In this case, the objective function did not attain zero so that the margins are not exactly matched in each sample.

Table 6
Solution to Example 3

s				$p(s)$
$i = 1$		$i = 2$		
1	0	0	1	0.5
0	0	0	0	
0	0	0	0	0.5
0	1	1	0	

4. COMPARISON OF MSE

In this section the mean squared error (MSE) of the proposed design with estimator \bar{y} will be compared with the MSE of the design of Bryant *et al.* (1960) with either of the two estimators they propose, namely \bar{y}_U and \bar{y}_B , where the U and B subscripts indicate that the first estimator is unbiased and the second is not. Let the cells be denoted c (ij in the two-way case), let k (and where necessary l) denote a unit within a cell, and suppress the s in $n_c(s)$ for simplicity of notation. The inclusion probability of any unit k in cell c is

$$\pi_{ck} = E[n_c] / N_c = E[n_c] / (NP_c)$$

(4.1)

and the joint inclusion probability of unit k in cell c and unit k' in cell c' is

$$\pi_{ckc'k'} = \begin{cases} \frac{E[n_c(n_c-1)]}{N_c(N_c-1)} & \text{if } c = c' \\ \frac{E[n_c n_{c'}]}{N_c N_{c'}} & \text{if } c \neq c'. \end{cases}$$

(4.2)

For large N this is approximately

$$\pi_{ckc'k'} \doteq \frac{E(n_c n_{c'})}{N^2 P_c P_{c'}} - \frac{E(n_c)}{N^2 P_c^2} I_{[c=c']},$$

(4.3)

where

$$I_{[c=c']} = \begin{cases} 1 & \text{if } c = c' \\ 0 & \text{if } c \neq c'. \end{cases}$$

The expectations will differ for our design compared to the Bryant *et al.* design and thus the π_{ck} and $\pi_{ckc'k'}$ will differ. Keeping this in mind we can obtain the variance of \bar{y} , \bar{y}_U and \bar{y}_B in a generalized form in terms of the π_{ck}

and $\pi_{ckc'k'}$ values and thus have some basis for comparison. To do this, let us consider an estimator of the form $\bar{z} = \sum_c \sum_k w_c y_{ck} / n$, where the w_c values are fixed known constants independent of k . If we restrict to the case where $n_{i\cdot} = nP_{i\cdot}$ and $n_{\cdot j} = nP_{\cdot j}$, that is, integer marginal requirements, then both of the estimators given in Bryant *et al.* as well as our estimator are of this form. We will assume this to be the case in the sequel. Replacing the subscript c with ij for two-way stratification, \bar{y}_U and \bar{y}_B are of the same form as \bar{z} with $w_c = w_{ij} = G_{ij} = P_{ij} / (P_{i\cdot} P_{\cdot j})$ and $w_c = w_{ij} = 1$, respectively. The estimator \bar{y} is also of the form \bar{z} with $w_c = w_{ij} = 1$.

We can now obtain a general form for the variance of \bar{z} keeping in mind that the π_{ck} and $\pi_{ckc'k'}$ values will differ for the Bryant *et al.* design and our design:

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \sum_{c'} \sum_k \sum_{k'} (\pi_{ck} \pi_{c'k'} - \pi_{ckc'k'}) (w_c y_{ck} - w_{c'} y_{c'k'})^2. \quad (4.4)$$

Using (4.1) and (4.3) this becomes

$$\begin{aligned} V(\bar{z}) &= \frac{1}{2n^2} \sum_c \frac{w_c^2 E(n_c)}{N^2 P_c^2} \sum_k \sum_{k'} (y_{ck} - y_{ck'})^2 \\ &\quad - \frac{1}{2n^2} \sum_c \sum_{c'} \frac{\text{Cov}(n_c, n_{c'})}{N^2 P_c P_{c'}} \sum_k \sum_{k'} (w_c y_{ck} - w_{c'} y_{c'k'})^2. \end{aligned} \quad (4.5)$$

Noting that

$$\sum_k \sum_l (y_{ck} - y_{cl})^2 = 2N^2 P_c^2 S_c^2$$

and

$$\begin{aligned} \sum_k \sum_{k'} (w_c y_{ck} - w_{c'} y_{c'k'})^2 &= N^2 P_c P_{c'} \\ &\quad [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 + (w_c \bar{Y}_c - w_{c'} \bar{Y}_{c'})^2], \end{aligned}$$

where S_c^2 refers to the population variance of cell c , (4.5) reduces to

$$\begin{aligned} V(\bar{z}) &= \frac{1}{n^2} \sum_c w_c^2 E(n_c) S_c^2 \\ &\quad - \frac{1}{2n^2} \sum_c \sum_{c'} \text{Cov}(n_c, n_{c'}) [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 \\ &\quad + (w_c \bar{Y}_c - w_{c'} \bar{Y}_{c'})^2] \\ &= v_1 + v_2, \quad \text{say}. \end{aligned} \quad (4.6)$$

The first term v_1 may be interpreted as the usual stratified variance for fixed sample sizes $E(n_c)$ within the two-way 'strata' (of course in our case the $E(n_c)$ will generally not be integers). The second term v_2 may be interpreted as the increase in variance arising from the variability of the n_c and the correlation between them. We discuss this further at the end of this section. We now revert to the notation $c = ij$ and compare the variances for two-way stratification.

First let us consider v_1 in (4.6). For the Bryant *et al.* method $E(n_{ij}) = nP_{i\cdot} P_{\cdot j}$, $\bar{y}_U = \sum_i \sum_j \sum_k G_{ij} y_{ijk} / n$, $G_{ij} = P_{ij} / (P_{i\cdot} P_{\cdot j})$ and $\bar{y}_B = \sum_i \sum_j \sum_k y_{ijk} / n$.

Thus

$$v_1(\bar{y}_U) = \sum_i \sum_j P_{ij} G_{ij} S_{ij}^2 / n,$$

(this is the same as the first term of equation (12) in Bryant *et al.*) and

$$v_1(\bar{y}_B) = \sum_i \sum_j P_{i\cdot} P_{\cdot j} S_{ij}^2 / n.$$

In the case of our approach $E(n_{ij}) = nP_{ij}$ and $\bar{y} = \sum_i \sum_j \sum_k y_{ijk} / n$ so that

$$v_1(\bar{y}) = \sum_i \sum_j P_{ij} S_{ij}^2 / n.$$

Next let us consider v_2 . It is not difficult to show that for both the Bryant *et al.* method and our approach (see Appendix)

$$\sum_i \text{Cov}(n_{ij}, n_{i'j'}) = \sum_j \text{Cov}(n_{ij}, n_{i'j'}) = 0. \quad (4.7)$$

Using this and replacing c and c' by ij and $i'j'$, respectively, in v_2 given in (4.6), it follows that v_2 reduces to

$$v_2 = \frac{1}{n^2} \sum_i \sum_j \sum_{i'} \sum_{j'} \text{Cov}(n_{ij}, n_{i'j'}) w_{ij} w_{i'j'} \bar{Y}_{ij} \bar{Y}_{i'j'}.$$

Replacing w_{ij} with G_{ij} we get $v_2(\bar{y}_U)$, and using simple algebra one can show that this is the same as term 2 of equation (12) in Bryant *et al.* Replacing w_{ij} with 1 gives the form of $V(\bar{y}_B)$ and of $V(\bar{y})$, noting that the $\text{Cov}(n_{ij}, n_{i'j'})$ will not be the same for both. So we see that v_2 depends only on the cell means while v_1 depends only on the within cell variances.

Finally, we should note that

$$\text{bias}(\bar{y}_B) = - \sum_i \sum_j (P_{ij} - P_{i\cdot} P_{\cdot j}) \bar{Y}_{ij}, \quad (4.8)$$

since to compare the three estimators the mean square error (MSE) will be the relevant measure, and this bias will contribute to $\text{MSE}(\bar{y}_B)$.

Combining the expressions for v_1 , v_2 and bias(\bar{y}_B) above permits an analytical comparison of the MSE of the proposed approach with that of the approach of Bryant *et al.* (1960) using either \bar{y}_U and \bar{y}_B . It is difficult, however, to make general statements about the relative performance of the different strategies and so we now consider introducing some model assumptions in order to approximate the different components of the MSE expressions, in some specific settings. We first consider the additive model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

where y_{ijk} is the k -th observation in the ij -th cell, α_i and β_j are fixed effects and ϵ_{ijk} are independent errors with zero mean and common variance σ^2 . Then $E_m(S_{ij}^2) = \sigma^2$ and $E_m(\bar{Y}_{ij}\bar{Y}_{i'j'}) = (\mu + \alpha_i + \beta_j)(\mu + \alpha_{i'} + \beta_{j'})$. Thus the model-expected design-variance is given by replacing S_{ij}^2 by σ^2 and \bar{Y}_{ij} by $\mu + \alpha_i + \beta_j$ in the formulas for v_1 and v_2 for the various estimators. In this case, $v_2(\bar{y}_B) = 0$. This point was realized by Bryant *et al.* when comparing \bar{y}_U and \bar{y}_B . The bias term will be zero in this case unless there was rounding on the margins, that is bias(\bar{y}_B) = 0 provided $n_{i\cdot} = nP_{i\cdot}$ and $n_{\cdot j} = nP_{\cdot j}$ as is the case in their example. This easily follows from (4.8) and

$$\sum_i (P_{ij} - P_{i\cdot}P_{\cdot j}) = \sum_j (P_{ij} - P_{i\cdot}P_{\cdot j}) = 0.$$

This was also shown by Bryant *et al.* p. 119 equation (47). Using (4.7), it is easily shown that $v_2(\bar{y}) = 0$ as well. This combined with the unbiasedness of \bar{y} and the fact that $v_1(\bar{y}_B) = v_1(\bar{y}) = \sigma^2/n$ in this case implies that for this situation $\text{MSE}(\bar{y}_B) = \text{MSE}(\bar{y})$, that is the proposed procedure has the same MSE as the procedure of Bryant *et al.* using the biased estimator. We demonstrate in the sequel that even when this additive model is applicable ($\gamma = 0$ below), $v_2(\bar{y}_U)$ may be large while $v_1(\bar{y}_U) > v_1(\bar{y})$.

To compare the estimators further, let us consider the situation of Example 1. The above derivations allow us to obtain the MSE's of the three estimators for this example provided we have the S_{ij} 's, the \bar{Y}_{ij} 's and can calculate the Cov(n_{ij} , $n_{i'j'}$) for the Bryant *et al.* method as well as for our approach. The covariances for the

Bryant *et al.* method are given in their paper in terms of the P_{ij} 's, while the covariances for our approach can be obtained from the solution in Table 3. We will consider non-additive departures from the above model, namely

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j + \epsilon_{ijk},$$

for various values of γ . For simplicity of presentation, let $\mu = 1$, $\alpha_i = i - 3$, $\beta_j = j - 2$ (note in fact that the MSE of each strategy is invariant to the choice of μ). Thus the model-expected design-variance is given by replacing S_{ij}^2 by 1 and \bar{Y}_{ij} by $1 + (i - 3) + (j - 2) + \gamma(i - 3)(j - 2)$ in the formulas for v_1 and v_2 for the various estimators. Table 7 gives the resulting v_1 , v_2 , and MSE values for the three estimators (as well as the bias squared term for \bar{y}_B), for various values of γ . From Table 7, it can be seen that for an additive model, $\gamma = 0$, \bar{y}_B and \bar{y} perform equally well, while \bar{y}_U is inferior. As the model becomes more non-additive, and $|\gamma|$ increases, the two estimators for the Bryant *et al.* strategy tend to perform similarly, both with MSE becoming increasingly greater than that of the proposed strategy. This pattern is primarily due to the v_2 component of the MSE of the three estimators. The bias term of \bar{y}_B is of lesser importance, although it may be more important for larger n .

The greater increase in v_2 as $|\gamma|$ increases for the Bryant *et al.* design appears to reflect the greater variability of each n_{ij} for this design. It should be noted that it would have been possible to reduce this variability somewhat by applying a variant of the Bryant *et al.* method instead to Table 2, as was done for the proposed method, though one would need to derive adjusted G_{ij} weights for \bar{y}_U and it would be difficult to handle the 0.0 cell entries in Table 2. However, even if this were accomplished, the \bar{n}_{ij} for this design may still take values other than just 0 and 1; for example n_{42} could take values 0, 1, or 2. This inflated n_c variability is inherent in the Bryant *et al.* method. For example, suppose $n_{1\cdot} = n_{\cdot 1} = 5$. Then using the Bryant *et al.* method, n_{11} can take values 0, 1, 2, 3, 4, or 5, while with the proposed method it can take only values $[nP_{11}]$ or $[nP_{11}] + 1$. If $nP_{11} < 1$, the technique used to go from Table 1 to Table 2 will not improve matters.

Table 7
Comparison of MSE for Three Estimators

γ	Bryant, Hartley, Jessen Design							Proposed Design		
	\bar{y}_U			\bar{y}_B				\bar{y}		
	v_1	v_2	MSE	v_1	v_2	Bias ²	MSE	v_1	v_2	MSE
0	.125	.105	.230	.100	.000	.000	.100	.100	.000	.100
$\pm .5$.125	.063	.188	.100	.033	.002	.135	.100	.018	.118
± 1	.125	.105	.230	.100	.131	.008	.239	.100	.071	.171
± 2	.125	.440	.565	.100	.523	.032	.655	.100	.284	.384
± 3	.125	1.111	1.236	.100	1.176	.073	1.349	.100	.638	.738

5. VARIANCE ESTIMATION

In this section, we will consider variance estimation for our proposed method. Using (4.1) and recalling constraint (2.1), it is clear that

$$\pi_{ck} = E[n_c(s)/N_c] = n/N.$$

The joint inclusion probability of two units k, k' in the same cell c is

$$\pi_{ck,ck'} = E[n_c(s)\{n_c(s) - 1\}/\{N_c(N_c - 1)\}].$$

Suppose $n_c(s) = I_c + \tilde{n}_c(s)$ when I_c is the fixed integer $[nP_c]$ and $\tilde{n}_c(s) = 0$ or 1 .

If $nP_c \leq 1$ then $I_c = 0$ and $\pi_{ck,ck'} = 0$. Hence a necessary condition for unbiased variance estimation to be possible is that $nP_c > 1$ for all cells c . On the other hand if this condition holds then $n_c(s) \geq 1$ for all c and hence the probability of inclusion of any pair of units in different cells is also always positive. Hence this condition is necessary and sufficient for unbiased variance estimation to be possible.

When this condition holds we obtain

$$\pi_{ck,ck'} = I_c(I_c + 2r_c - 1)/[N_c(N_c - 1)] = A_c,$$

say, where $r_c = E[\tilde{n}_c(s)] = nP_c - I_c$.

The joint inclusion probability for pairs of units in different cells c and c' are

$$\begin{aligned} \pi_{ck,ck'} &= E[n_c(s)n_{c'}(s)/(N_cN_{c'})] \\ &= [I_cI_{c'} + r_{c'}I_c + r_cI_{c'} + r_{cc'}]/(N_cN_{c'}) = B_{cc'}, \end{aligned} \quad (5.1)$$

say where $r_{cc'} = E[\tilde{n}_c(s)\tilde{n}_{c'}(s)]$.

Hence an unbiased estimator of $V(\bar{y}(s))$ of Sen-Yates-Grundy form may be constructed in the usual way.

In practice, however, we wish to consider situations where $nP_c \leq 1$ for some c . In this case one assumption we might make following Bryant *et al.* (1960, Sect. 7) in order to derive a variance estimator is that the population variance of Y is constant within each cell c , say S^2 .

Let us first obtain the variance of $\bar{y}(s)$ in the general case

$$\begin{aligned} V(\bar{y}(s)) &= \frac{1}{2n^2} \sum_c \sum_{k \neq k'} \sum \left(\frac{n^2}{N^2} - A_c \right) (y_{ck} - y_{ck'})^2 \\ &\quad + \frac{1}{2n^2} \sum_{c \neq c'} \sum_{k, k'} \sum \left(\frac{n^2}{N^2} - B_{cc'} \right) (y_{ck} - y_{c'k'})^2. \end{aligned}$$

Now providing $B_{cc'} > 0 \forall c, c'$ we may estimate the second term unbiasedly by

$$\frac{1}{2n^2} \sum_A \sum_{k=1}^{n_c(s)} \sum_{k'=1}^{n_{c'}(s)} \left(\frac{\frac{n^2}{N^2} - B_{cc'}}{B_{cc'}} \right) (y_{ck} - y_{c'k'})^2,$$

where $A = \{c, c': n_c(s) \geq 1, n_{c'}(s) \geq 1, c \neq c'\}$.

The first term can be written as

$$\frac{1}{2n^2} \sum_c \left(\frac{n^2}{N^2} - A_c \right) 2N_c^2 S^2.$$

For any c s.t. $n_c(s) \geq 2$

$$E \left(\sum_{\substack{k=1 \\ k \neq k'}}^{n_c(s)} \sum_{k'=1}^{n_c(s)} \frac{(y_{ck} - y_{ck'})^2}{2n_c(s)\{n_c(s) - 1\}} \middle| n_c(s) \right) = S^2.$$

Thus provided at least one $n_c(s)$ is ≥ 2 an unbiased estimator of the first term is

$$\begin{aligned} \frac{1}{2n^2 D} \sum_{\{c: n_c(s) \geq 2\}} \left(\frac{n^2}{N^2} - A_c \right) 2N_c^2 \sum_{k=1}^{n_c(s)} \sum_{k'=1}^{n_c(s)} \\ \frac{(y_{ck} - y_{ck'})^2}{2n_c(s)\{n_c(s) - 1\}} \end{aligned}$$

where D = the number of cells, c , such that $n_c(s) \geq 2$.

The above requires $B_{cc'} > 0$. If

$$I_c = I_{c'} = 0,$$

by (5.1), we need

$$r_{cc'} = \sum \tilde{n}_c(s)\tilde{n}_{c'}(s)p(s) > 0, \quad (5.2)$$

which is linear in $p(s)$. The constraint (5.2) can be handled in linear programming if desired. There will be such a constraint for each pair c, c' s.t. $I_c = I_{c'} = 0$.

6. CONCLUDING REMARKS

We have proposed a linear programming approach to multi-way stratification, applying ideas of Rao and Nigam (1990, 1992). The approach is simple in conception and is very flexible in allowing for a range of different objectives via the loss function $w(s)$, as well as in permitting

a variety of constraints such as that the joint inclusion probabilities of all cells be positive. The main practical constraint on the procedure is that it may rapidly become computationally expensive if not impossible as the number of cells in the multi-way classification increases. Some ideas on how to reduce the amount of computation have been considered. Further research on this question would be useful. For cases where the computational demands are prohibitive, the method of Causey *et al.* (1985) remains an alternative.

ACKNOWLEDGEMENTS

The authors wish to express their thanks to Wesley Yung for his computer programming assistance and J.N.K. Rao for helpful comments in the early stages of discussion. R.R. Sitter is supported by the Natural Sciences and Engineering Council of Canada. Our thanks are due to the referees and the Associate Editor for constructive suggestions and comments.

APPENDIX

Proof of (4.7) for Proposed Method

Note that

$$\begin{aligned} \text{Cov}(n_{ij}(s), n_{i'j'}(s)) &= E(n_{ij}(s)n_{i'j'}(s)) \\ &\quad - E(n_{ij}(s))E(n_{i'j'}(s)). \end{aligned}$$

Equation (2.1) states that $E(n_{ij}(s)) = nP_{ij}$. By definition

$$E(n_{ij}(s)n_{i'j'}(s)) = \sum_s n_{ij}(s)n_{i'j'}(s)p(s).$$

Thus

$$\sum_j E(n_{ij}(s))E(n_{i'j'}(s)) = n^2 P_{i'j'} \sum_j P_{ij} = n^2 P_{i'j'} P_{i\cdot}, \quad (7.1)$$

and

$$\begin{aligned} \sum_j E(n_{ij}(s)n_{i'j'}(s)) &= \sum_j \sum_s n_{ij}(s)n_{i'j'}(s)p(s) \\ &= \sum_s p(s)n_{i'j'}(s) \sum_j n_{ij}(s). \end{aligned} \quad (7.2)$$

Assume that the solution to the linear optimization problem (2.2) equals zero, where $w(s)$ is given in (2.9). In this case, $\sum_j n_{ij}(s) = n_{i\cdot}(s) = nP_{i\cdot}$ and (7.2) implies

$$\begin{aligned} \sum_j E(n_{ij}(s)n_{i'j'}(s)) &= \sum_s p(s)n_{i'j'}(s)nP_{i\cdot} \\ &= nP_{i\cdot} \sum_s n_{i'j'}(s)p(s) \\ &= nP_{i\cdot} E(n_{i'j'}(s)) = nP_{i\cdot} nP_{i'j'}. \end{aligned} \quad (7.3)$$

Equations (7.1) and (7.3) together imply $\sum_j \text{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0$. It can be similarly shown that

$$\begin{aligned} \sum_i \text{Cov}(n_{ij}(s), n_{i'j'}(s)) &= \sum_{i'} \text{Cov}(n_{ij}(s), n_{i'j'}(s)) \\ &= \sum_{j'} \text{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0. \end{aligned}$$

REFERENCES

- BRYANT, E.C., HARTLEY, H.O., and JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.
- CAUSEY, B.D., COX, L.H., and ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- GOODMAN, R., and KISH, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.
- RAO, J.N.K., and NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.
- RAO, J.N.K., and NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *International Statistical Review*, 60, 89-98.
- WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey

WAYNE A. FULLER, MARIE M. LOUGHIN and HAROLD D. BAKER¹

ABSTRACT

A regression weight generation procedure is applied to the 1987-1988 Nationwide Food Consumption Survey of the U.S. Department of Agriculture. Regression estimation was used because of the large nonresponse in the survey. The regression weights are generalized least squares weights modified so that all weights are positive and so that large weights are smaller than the least squares weights. It is demonstrated that the regression estimator has the potential for large reductions in mean square error relative to the simple direct estimator in the presence of nonresponse.

KEY WORDS: Non-negative weights; Consistency.

1. INTRODUCTION

In many sampling situations, the population means of auxiliary variables are known, but the particular values of the variables for individual elements are not used in the sample selection. Although the information is not used in the sampling design, it may be highly desirable to incorporate the information about population means into the estimation procedure. Common estimation procedures utilizing auxiliary information are ratio estimation, post-stratification, regression estimation, and raking. Regression estimation is the most general procedure in that the regression method can handle multiple auxiliary variables, continuous auxiliary variables, and discrete auxiliary variables. Post-stratification can be considered a special case of regression estimation in which the regression variables are indicator variables for the post strata. The raking procedure, also known as iterative proportional fitting, is restricted to auxiliary information in the form of discrete categories. Deming and Stephan (1940), Stephan (1942), El-Badry and Stephan (1955), Ireland and Kulblack (1968), Darroch and Ratcliff (1972), Brackstone and Rao (1979), and Oh and Scheuren (1987) are references on raking.

Early applications of regression estimation are Watson (1937), Cochran (1942) and Jessen (1942). Cochran (1977, Ch. 7) contains the basic theory. Regression estimation for survey samples has been discussed by numerous authors, including Mickey (1959), Fuller (1975), Royall and Cumberland (1981), Isaki and Fuller (1982), Wright (1983), Luery (1986), Alexander (1987), Bethlehem and Keller (1987), Copeland, Pritzeimer, and Hoy (1987), Lemaître and Dufour (1987), Särndal, Swensson and Wretman (1989), Deville and Särndal (1992), Zieschang (1990), and Rao (1992).

In much of the cited literature, regression estimation is described as a procedure for reducing variance in probability samples. In practice, one of the motivations for regression estimation is the potential for reducing bias associated with selective nonresponse. See, for example, Little and Rubin (1987, p. 55) for the special case of adjustment cells, and Bethlehem (1988) for the generalized regression estimator.

Nonresponse prompted the use of regression estimation in our application and we discuss regression estimation in the response adjustment context in Section 3. The standard regression estimator and the modified procedure that produces positive weights are introduced in Section 2. Application of the regression weighting procedure to the Nationwide Food Consumption Survey is described in Section 4.

2. REGRESSION ESTIMATOR

To introduce the regression estimator used in our study, assume that a sample containing n units has been selected and that the probability of selecting unit i is π_i . For our purposes, it is sufficient for π_i to be proportional to the selection probabilities. The sample might be a two-stage stratified sample, and the unit can be either the primary sampling unit or the observation unit. In our application, the unit is the observation unit. Assume that a k -dimensional vector of population means, denoted by $\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$ is known, that the vector $(y_i, x_{i1}, x_{i2}, \dots, x_{ik})$ is observed for every unit in the sample and that an estimator of the mean of y is desired. We assume that the first element of x_i is one for all i . Hence, the first element of \bar{X} is also one. The vector $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ is sometimes

¹ Wayne A. Fuller, Marie M. Loughin and Harold D. Baker, Iowa State University.

called the vector of control variables. A regression estimator of the mean of y is

$$\bar{y}_r = \bar{X}\hat{\beta}, \quad (2.1)$$

where

$$\hat{\beta} = \left(\sum_{i=1}^n x_i' \pi_i^{-1} x_i \right)^{-1} \sum_{i=1}^n x_i' \pi_i^{-1} y_i, \quad (2.2)$$

and we have assumed $\sum x_i' \pi_i^{-1} x_i$ to be nonsingular. This definition of the regression estimator follows Mickey (1959) who suggested restricting the term regression estimator to estimators that are location and scale invariant. The estimator (2.1) can also be written as

$$\bar{y}_r = \sum_{i=1}^n w_i y_i, \quad (2.3)$$

where

$$w_i = \bar{X} \left(\sum_{i=1}^n x_i' \pi_i^{-1} x_i \right)^{-1} x_i' \pi_i^{-1}, \quad (2.4)$$

and the weights have the property,

$$\sum_{i=1}^n w_i x_i = \bar{X}. \quad (2.5)$$

The weights of expression (2.4) are relatively easy to compute, and once computed, can be used for the estimation of any y -variable. If the vector x_j is replaced by the vector

$$(1, z_j) = (1, x_{j2} - \bar{X}_2, x_{j3} - \bar{X}_3, \dots, x_{jk} - \bar{X}_k), \quad (2.6)$$

the estimator can be written in the form

$$\bar{y}_r = \bar{y}_\pi + (\bar{Z} - \bar{z}_\pi) \hat{\beta}_z = \bar{y}_\pi - \bar{z}_\pi \hat{\beta}, \quad (2.7)$$

where $\bar{Z} = 0$ is the population mean of z_j , $\bar{z}_\pi = \bar{x}_\pi - \bar{X}$,

$$(\bar{y}_\pi, \bar{z}_\pi) = \left(\sum_{i=1}^n \pi_i^{-1} \right)^{-1} \sum_{i=1}^n \pi_i^{-1} (y_i, z_i)$$

and

$$\hat{\beta}_z = \left[\sum_{j=1}^n (z_j - \bar{z}_\pi)' \pi_i^{-1} (z_j - \bar{z}_\pi) \right]^{-1} \sum_{j=1}^n (z_j - \bar{z}_\pi)' \pi_i^{-1} y_j.$$

In the form (2.7), \bar{y}_r is the intercept in the regression of y on z . Thus, the theory given by Fuller (1975) for regression coefficients is applicable to the regression estimator of the mean. If the population total of units is known and denoted by N , the estimated population total is $N\bar{y}_r$.

By defining a sequence of populations and samples, it is possible to show that the estimator (2.1) is a consistent estimator of the mean of y . See, for example, Fuller (1975). The estimator of the variance of the regression estimator is a function of the joint probabilities. Consider a stratified two-stage sample and replace our single subscript i with the triple ℓjt . Then, omitting the finite correction term, a variance estimator is

$$\hat{V}\{\bar{y}_r\} = (n - k)^{-1} n \sum_{\ell=1}^L (n_\ell - 1)^{-1} n_\ell \sum_{j=1}^{n_\ell} (d_{\ell j} - d_{\ell..})^2, \quad (2.8)$$

where

$$d_{\ell j} = \sum_{t=1}^{m_{\ell j}} w_{\ell jt} (y_{\ell jt} - x_{\ell jt} \hat{\beta}),$$

$$d_{\ell..} = n_\ell^{-1} \sum_{j=1}^{n_\ell} d_{\ell j},$$

n_ℓ is the number of sample primary sampling units in stratum ℓ , $m_{\ell j}$ is the number of sample elements in primary sampling unit j of stratum ℓ , $\hat{\beta}$ is the vector of coefficients defined in (2.2), n is the total number of elements in the sample, and $w_{\ell jt}$ is the weight for element t in primary sampling unit j of stratum ℓ . The factor $n - k$ is used by analogy to the divisor for the unbiased estimator of the error variance in ordinary regression. When the vector of control variables is coded as in (2.6), the estimator (2.8) is the estimated variance of the first element of $\hat{\beta}$, the estimated intercept. The estimator (2.8) was suggested in Hidioglou, Fuller and Hickman (1976) and the consistency of the estimator was established by Fuller (1975). Also see Särndal, Swensson and Wretman (1989).

The estimators, constructed with weights (2.4), have good large sample properties. However, they may have undesirable behavior in small samples. Because the weights are linear functions of x_i , it is possible for some of the weights to be negative. Negative weights make it possible for estimates of positive parameters to be negative. Early research on methods of constructing nonnegative regression weights was conducted by Husain (1969). Huang (1978) designed a computer program to produce nonnegative regression weights. Huang and Fuller (1978) described the weight generation procedure and showed

that the large sample distribution of the modified estimator is the same as that of the ordinary regression estimator. Also see Goebel (1976).

The computer algorithm of Huang (1978) is an iterative procedure based upon the ideas of generalized least squares. The goal of the Huang algorithm is a set of weights w_i , $i = 1, 2, \dots, n$, satisfying (2.5) that do not differ greatly from the initial weights, where difference is a function of the initial weight. The Huang algorithm attempts to compute weights w_i satisfying

$$(1 - M) \max_{1 \leq i \leq n} w_i \pi_i^{-1} \leq (1 + M) \min_{1 \leq i \leq n} w_i \pi_i^{-1},$$

where the parameter M , $0 < M \leq 1$, is specified by the user and is generally chosen in the interval $[0.8, 1.0]$. If the first round regression weights defined by (2.4) do not satisfy the requirements, a second round of regression weights is computed. The second round weights are weighted regression weights in which a control factor is assigned to each observation. Small control factors are assigned to observations with large or small first round weights. Relatively large control factors are assigned to observations with first round weights close to π_i^{-1} . The second round regression weights are checked and if they fail to satisfy the criteria, the control factors are modified, and so on. The algorithm is given in the Appendix.

The control weighting used in the Huang algorithm has much in common with bounded-influence and robust regression methods. That is, in the final estimator, the contribution to the estimation of the slope vector is reduced for observations that are far from the mean. See Hampel (1978), Krasker (1980), and Mallows (1983). Recent research on this type of estimator for survey samples is that of Deville and Särndal (1992), Akkerboom, Sikkels, and van Herk (1991), Hulliger (1993) and Singh (1993).

It is not always possible to construct weights meeting the criteria and also satisfying (2.5). For example, if all of the observations on x_{i2} exceed the mean, there is no set of positive weights summing to one that also satisfy $\sum_{i=1}^n x_{i2} w_i = \bar{X}_2$. Therefore, the weight generation program will terminate if weights meeting the specified criteria cannot be constructed after a specified number of iterations.

In some situations it is desirable to restrict the weights to the nonnegative integers. This is true when estimates of totals are being constructed and the population contains well defined units, such as people. Nonnegative integer weights then provide more comfortable estimates, in that the estimates are physically attainable. Integer weights can be constructed so that no rounding is necessary when building tables. With such integer weights, all multiple way tables will automatically be internally consistent.

The Huang program contains an option to round the real weights to integer weights in a manner that maintains

the sum of the weights. After rounding, the equalities (2.5) will generally no longer hold exactly. We have found that by iterating the Huang algorithm using the first-round integer weights as initial weights, integer weights can be constructed such that there is a very modest deviation from equality for expression (2.5). Cox (1987), Cox and Ernst (1982), and Fagan, Greenberg and Hemmig (1988) discuss rounding.

3. REGRESSION ESTIMATION WITH NONRESPONSE

The early theoretical developments for regression estimation assumed the sample to be a probability sample from the population. However, it has long been recognized that regression estimation can be used to reduce the bias that arises from imperfections in the data collection procedure. The most obvious of these imperfections is nonresponse. In all large samples of human subjects, some of the subjects fail to provide information. If the nonrespondents differ from the respondents, direct estimates constructed from the respondents will be biased. Given auxiliary information, regression estimation provides a method of reducing the bias. The degree to which the bias is reduced depends upon the relationship between the control variables, the variables of interest, and the response probabilities. See Little and Rubin (1987) for a general discussion of these issues.

Let π_i^* denote the inclusion probability equal to the product of π_i and the conditional probability of observing the unit given that the unit is selected. Then

$$E \left\{ \sum_{i=1}^n x_i' \pi_i^{-1} x_i \mid \xi_N \right\} = \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* x_i \quad (3.1)$$

and

$$E \left\{ \sum_{i=1}^n x_i' \pi_i^{-1} y_i \mid \xi_N \right\} = \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* y_i, \quad (3.2)$$

where the expectations are conditional on the given finite population ξ_N , and n is the realized sample size. In the case of nonresponse, the ratio $p_i = \pi_i^* \pi_i^{-1}$ is the response probability for individual i . Therefore, under conditions such as those used by Fuller (1975),

$$\text{plim}_{\substack{n \rightarrow \infty \\ N \rightarrow \infty}} (\hat{\beta} - \gamma) = 0, \quad (3.3)$$

where $\hat{\beta}$ is defined in (2.2) and

$$\gamma = \left(\sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* x_i \right)^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* y_i. \quad (3.4)$$

Then

$$\bar{Y} = \bar{X}\gamma + \bar{A}, \quad (3.5)$$

where $\bar{A} = N^{-1} \sum_{i=1}^N a_i$ and $a_i = y_i - x_i\gamma$. Thus, the regression estimator (2.1) will be a consistent estimator of \bar{Y} if $\text{plim}_{N \rightarrow \infty} \bar{A} = 0$. The probability limit of \bar{A} will be zero if the finite population is a random sample from an infinite population in which the linear model

$$y_i = x_i\beta + e_i, \quad E\{e_i\} = 0$$

holds for all i .

The mean \bar{A} is zero when $\pi_i^* = \pi_i$ for all i and an element of x_i is one for all i because then

$$\gamma = \beta = \left(\sum_{i=1}^N x_i' x_i \right)^{-1} \sum_{i=1}^N x_i' y_i \quad (3.6)$$

and $\sum_{i=1}^N (y_i - x_i\beta) = 0$. A sufficient condition for \bar{A} to be zero is the existence of a row vector c such that

$$cx_i' = \pi_i^{*-1} \pi_i = p_i^{-1}, \quad (3.7)$$

for $i = 1, 2, \dots, N$. Thus, if the ratio of nominal probabilities to true probabilities is a linear function of the control variables, the regression estimator is a consistent estimator of the mean of y , where the limit is for sequences as defined in Fuller (1975). One way in which (3.7) can be satisfied is for the elements of x_i to be dummy variables that define subgroups and for the response probabilities to be constant in each subgroup. This situation is sometimes described by saying that elements are missing at random in each subgroup. We take the assumption that $\bar{A} = 0$ as our working assumption in the empirical analysis.

In any regression problem, it is impossible to use the sample to verify some of the assumptions. For example, in ordinary least squares regression, the residuals $\hat{e}_i = y_i - x_i\hat{\beta}$ are uncorrelated with x_i and, hence, the residuals cannot be used to check the assumption that the true errors are uncorrelated with x . Thus, in a survey with nonresponse, one searches for control variables that are correlated with y and (or) that one believes are correlated with the response probabilities. But one cannot guarantee that all bias has been removed by regression estimation based on a particular set of control variables.

In practice, one can often identify x -variables that are correlated with the probability of response and (or) correlated with the y -variables. For example, in the 1987-1988 Nationwide Food Consumption Survey, the response rate was low among high-income households. Therefore, use of variables for household income in a regression estimator is expected to reduce the bias in the estimated mean for characteristics that are correlated with income.

The error in $\hat{\beta}$ as an estimator of γ can be approximated by

$$\hat{\beta} - \gamma \doteq G^{-1} T^{-1} \sum_{i=1}^n x_i' \pi_i^{-1} a_i,$$

where a_i is defined in (3.5),

$$T = \sum_{i=1}^N \pi_i^{-1} \pi_i^*$$

and

$$G = T^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* x_i.$$

Under reasonable assumptions

$$\hat{T} = \sum_{i=1}^n \pi_i^{-1}$$

and

$$\hat{G} = \hat{T}^{-1} \sum_{i=1}^n x_i' \pi_i^{-1} x_i$$

are consistent estimators of T and G . Thus, the variance of the regression estimator can be estimated by estimating the variance of $\sum_{i=1}^n x_i' \pi_i^{-1} a_i$. If we assume that the conditional probabilities of response in one primary sampling unit are independent of those in all other primary sampling units and that at least one observation unit is observed in each selected primary sampling unit, then (2.8) remains an appropriate estimator of the variance of the regression estimated mean of y .

The estimator of variance (2.8) also remains appropriate if the regression weights are constructed by a procedure other than (2.4). For example, let the weights be defined by

$$w_{gi} = \bar{X} \left[\sum_{i=1}^n x_i' \pi_i^{-1} g_i x_i \right]^{-1} x_i' \pi_i^{-1} g_i,$$

where the g_i are functions of the x_i . Assume

$$\text{plim} \hat{\beta}_g = \gamma_g,$$

where

$$\hat{\beta}_g = \left[\sum_{i=1}^n x_i' \pi_i g_i x_i \right]^{-1} \sum_{i=1}^n x_i' \pi_i^{-1} g_i y_i.$$

Also assume

$$\text{plim}_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N (y_i - x_i \gamma_g) = 0.$$

Then expression (2.8) with w_{gi} replacing $w_{\ell i}$ is a consistent estimator of the variance of the estimator. The estimator (2.8) will be used in our empirical analyses.

Formula (2.8) identifies the two effects of regression estimation on the variance of an estimated mean. The correlation effect reduces the variance of the estimated mean while the increase in the sum of squares of the weights increases the variance of the estimated mean. To understand these effects, consider a simple random sample. If the y variable is correlated with x , the correlation tends to reduce the variance of the regression estimator relative to that of the simple estimator because

$$E\{(y_i - x_i\beta)^2\} \leq E\{[y_i - E(y_i)]^2\}.$$

If the sample means of the control variables differ from the population means, then

$$\sum_{i=1}^n w_i^2 > n^{-1},$$

where n^{-1} is the sum of squares of the simple weights for a simple random sample.

When comparing the variance of the sample mean with the variance of the regression estimator, one should not forget that one of the reasons for using regression estimation in samples with nonresponse is to produce an estimator with less bias than that of the direct estimator. Thus, in the next section we compare an estimator of the mean square error of the simple estimator to an estimator of the variance of the regression estimator.

4. APPLICATION TO THE NATIONWIDE FOOD CONSUMPTION SURVEY

The 1987-1988 Nationwide Food Consumption Survey was conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. The original sample was a self-weighting stratified sample of area primary sampling units within the 48 conterminous states. Primary sampling units were divided into secondary units called area segments. Households within the sample segments were contacted by personal interview. The field operation was conducted during the period April 1987 through August 1988 by a contractor under contract to the Human Nutrition Information Service.

Approximately 37% of the housing units identified as occupied provided complete household food use information. The realized household sample contains 4,495 households. Because of the low response rate, the Human Nutrition Information Service decided to use regression weighting in the estimation. Population totals for all characteristics except urbanization were estimated by the Human Nutrition Information Service from the March 1987 Current Population Survey. See Bureau of the Census (1987). The population totals for urbanization classes were furnished by the contractor. In our analysis, we treat the estimated population totals as if they were known population totals.

Table 1
Sample and population characteristics of households

Characteristic	Category	Household Sample Frequency	Household Sample Percent	Population Percent
Season	Spring	1,828	40.7	25.0
	Summer	678	15.1	25.0
	Fall	717	16.0	25.0
	Winter	1,272	28.3	25.0
Region	Northeast	905	20.1	21.2
	Midwest	1,172	26.1	24.7
	South	1,567	34.9	34.4
	West	851	18.9	19.6
Urbanization	Central Cities	1,064	23.7	31.2
	Suburban	2,122	47.2	46.0
	Nonmetro	1,309	29.1	22.9
Household Income as % of Poverty	< 131%	1,041	23.2	20.0
	131-300%	1,564	34.8	32.2
	301-500%	1,108	24.6	25.9
	> 500%	782	17.4	21.8
Household Food Stamps	Yes	314	7.0	7.4
	No	4,181	93.0	92.6
Ownership of Domicile	Yes	2,998	66.7	64.1
	No	1,497	33.3	35.9
Race of Household Head	Black	519	11.5	11.1
	Nonblack	3,976	88.5	88.9
Age of Household Head	< 25	338	7.5	7.9
	25-39	1,588	35.3	36.1
	40-59	1,369	30.5	30.5
	60-69	660	14.7	13.0
	70+	540	12.0	12.6
Household Head Status	Both Male and Female	3,057	68.0	60.8
	Female Only	1,044	23.2	26.0
	Male Only	394	8.8	13.2
Female Head Worked	Yes	1,792	39.9	41.5
	No	2,703	60.1	58.5
Exactly One Adult in Household	Yes	1,211	26.9	29.7
	No	3,284	73.1	70.3
Exactly Two Adults in Household	Yes	2,616	58.2	54.2
	No	1,879	41.8	45.8
Presence of Child < 7 Years Old	Yes	1,009	22.4	20.1
	No	3,486	77.6	79.9
Presence of Child 7-17 Years Old	Yes	1,309	29.1	26.5
	No	3,186	70.9	73.5
Household Size	(Mean)		2.731	2.642
Household Size, Squared	(Mean)		9.546	9.125

Characteristics of the population and of the household sample are given in Table 1. The original sample was unbalanced with respect to time of interview with nearly 41% of the interviews in the spring quarter and about 16% of the interviews in each of the summer and fall quarters. Interviews for the spring and summer quarters were done in both 1987 and 1988.

The sample was also unbalanced with respect to urbanization. There was a lower fraction of central city households in the sample than in the population (24% versus 31%), and a higher fraction of nonmetropolitan households in the sample than in the population (29% versus 23%).

The fraction of high income households was smaller in the sample than in the population. The sample contained a higher fraction of households with both a male and female head than the population (68% versus 61%). A higher fraction of the sample than of the population consisted of households with children. The sample was only mildly unbalanced with respect to several other socio-demographic characteristics.

The characteristics listed in Table 1 are believed by the staff of the Human Nutrition Information Service to be related to food consumption behavior. Therefore, variables based on those characteristics were used in the regression weighting procedure. To implement the weight generation program, each of the categorical variables of Table 1 was converted to a set of indicator variables. For example, three variables were created for the characteristic, household income as a percent of poverty. These were

$$Z_{t1} = 1 \quad \text{if income} < 131\% \text{ for } t\text{-th household} \\ = 0 \quad \text{otherwise,}$$

$$Z_{t2} = 1 \quad \text{if income is } 131\text{--}300\% \text{ for } t\text{-th household} \\ = 0 \quad \text{otherwise,}$$

$$Z_{t3} = 1 \quad \text{if income is } 301\text{--}500\% \text{ for } t\text{-th household} \\ = 0 \quad \text{otherwise.}$$

Using this procedure, 25 indicator variables were created. In addition, household size and the square of household size were used as continuous variables.

The twenty-seven variables were used to generate regression weights using Huang's program. The parameter M of the weight generation program was set equal to 0.9 in the computation. The weights were rounded to integers, where each integer weight is a weight in thousands. The sum of the weights is 88,942, which is the number of households in the population in thousands. The average weight is 19.787, the smallest weight is 6, and the largest weight is 47. Thus, the largest weight is 2.38 times the average weight. The sum of squares of the weights is 2,317,930. The average weight squared and multiplied by the sample size is 1,759,884. Thus, if a variable has zero multiple correlation with the 27 variables, the variance of an estimate computed with the weights will be about 1.32 times the variance of the simple unweighted estimator.

Figure 1 shows the regression weights computed by the Huang algorithm plotted against the ordinary least squares weights. Because there are 4,495 households, many points are hidden. Both weights are standardized by dividing by the average weight. Thus, the average for weights coded in this manner is one. Because there are 27 control variables used in the construction, the Huang weights tend to form a swarm of points about an S-shaped function of the original weights. If there were only one control variable, the points would fall on an S-shaped curve. The original weights for observations to the left of zero were negative.

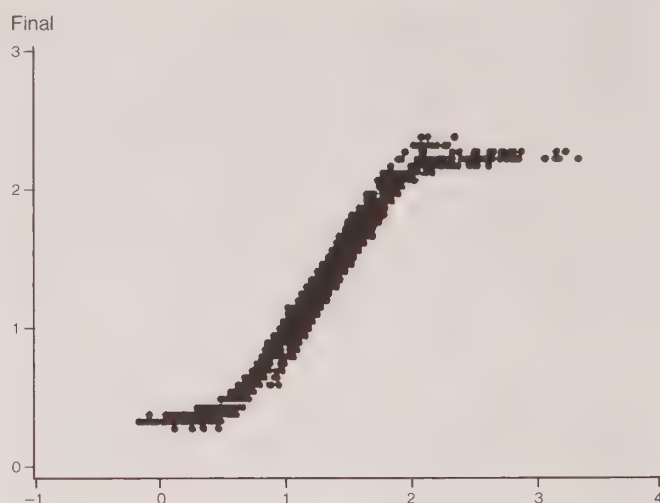


Figure 1. Plot of final weights against the ordinary least squares weights, both expressed relative to the average weight.

To compare estimates constructed with weights to unweighted estimates, we use the variables

Y_1 = adjusted total number of meals away from home (meals away),

Y_2 = total money value of food used at home (home food),

Y_3 = household size in 21-meal-equivalent persons (meal persons),

Y_4 = indicator to identify housekeeping households (housekeeping).

The household size in 21-meal-equivalent persons is the total adjusted meals eaten from household food supplies in the past 7 days divided by 21. "Meal persons" is the sum of two terms. The first term is the sum of the proportions of meals eaten at home in the interview week by each household member. The second term is the number of meals served to guests, boarders, and employees during the interview week, divided by 21. In other words:

$$\text{meal persons for } j\text{-th household} = \sum_i (h_{ij} + a_{ij})^{-1} h_{ij} + (21)^{-1} b_j,$$

where h_{ij} = meals eaten at home by the i -th individual in the j -th household during the interview week, a_{ij} = meals eaten away from home by the i -th individual in the j -th household during the interview week, b_j = number of meals eaten by nonhousehold members in the j -th household during the interview week.

The adjusted total number of meals bought and eaten away from home is the sum of the proportions of meals eaten away from home in the interview week by household members, multiplied by 21. In the notation used to define meal persons,

$$\text{meals away for } j\text{-th household} = 21 \sum_i (h_{ij} + a_{ij})^{-1} a_{ij}.$$

The total value of food used at home is the expenditures for purchased food plus the money value of home-produced food and food received free-of-cost that was used during the survey week. Expenditures for purchased food were based on prices reported as paid regardless of the time of purchase. Sales tax was excluded. Purchased food with unreported prices, food produced at home, food received as a gift, and food received instead of pay were valued at the average price per pound paid for comparable food by survey households in the same region and season.

A housekeeping household is a household with at least one person having ten or more adjusted meals from the household food supply during the seven days before the interview. Household food-use analyses generally consider only housekeeping households.

Table 2
Properties of alternative estimators

Variable	Un-weighted Mean	Weighted Mean	Difference	Relative Efficiency of Regression
Meals away				
Housekeeping	7.75 (0.22)	7.93 (0.17)	-0.18 (0.09)	2.52
Nonhousekeeping	18.31 (1.14)	18.12 (1.19)	0.19 (0.68)	0.92
All	8.27 (0.22)	8.57 (0.22)	-0.30 (0.12)	2.56
Home food				
Housekeeping	61.10 (1.14)	59.56 (0.98)	1.54 (0.41)	3.65
Nonhousekeeping	25.99 (1.25)	26.39 (1.46)	-0.40 (1.00)	0.73
All	59.37 (1.12)	57.49 (0.91)	1.88 (0.39)	5.60
Meal persons				
Housekeeping	2.42 (0.03)	2.33 (0.01)	0.09 (0.01)	89.00
Nonhousekeeping	0.51 (0.03)	0.49 (0.03)	0.02 (0.02)	1.00
All	2.33 (0.03)	2.22 (0.01)	0.11 (0.01)	129.00
Housekeeping (%)	95.06 (0.40)	93.77 (0.58)	1.29 (0.10)	5.30

The means of the variables computed using unweighted data are given in Table 2 in the column headed, "Un-weighted mean". Three means are given for meals away, home food, and meal persons. Two means are computed for the two subpopulations defined by the housekeeping variables. The third mean, designated by "all" in the table,

is the estimated mean for the entire population. The standard errors of the estimates are given in parentheses below the estimates. The estimates and standard errors for the unweighted estimates were computed in PC CARP. See Fuller *et al.* (1986). The computations accounted for the fact that the sample is an area stratified cluster sample.

Because the sample is a two-stage sample, the estimated variances are larger than the variance of a simple random sample containing the same number of households. The ratio of the variance for a sample estimate to the variance of a simple random sample containing the same number of individuals is called the design effect. The estimated design effect is about 2.5 for meals away and meal persons, is about 4.1 for home food, and is about 1.5 for housekeeping for the "all" means for the unweighted sample.

The column headed "Weighted mean" contains the estimates computed with the regression weights. The standard errors were computed in PC CARP using formula (2.8) with the regression weights replacing the π_i^{-1} . The variance calculation requires computing a regression for every y -variable. The estimated means for the subpopulations are ratios of regression estimators. The variances for the subpopulation means were computed by calculating the variances of the Taylor deviates for the ratio using formula (2.8). The standard errors for unweighted and weighted estimates are similar for meals away and home food. However, the standard errors for the regression estimate of the population mean of meal persons is about one third of the standard error of the unweighted estimate. The standard error of the regression estimator is smaller because meal persons is highly correlated with the household size variables used as controls in the regression procedure.

The estimated squared multiple correlation between the variables of the table and the 27 control variables is 0.29, 0.44, 0.82, and 0.12 for meals away, home food, meal persons, and housekeeping, respectively. If the sample means of the control variables were nearly equal to the population means, the standard error of the regression estimate of meals away would be about $(1 - 0.29)^{1/2} = 0.84$ times the standard error of the unweighted estimate. In fact, the estimated standard error of the regression is about 0.97 times the standard error of the unweighted estimate. The difference is due to the fact that $\sum_{i=1}^n w_i^2$ is considerably bigger than n^{-1} because the sample is unbalanced on a number of items. Note that

$$0.97 \doteq [(0.71)(1.32)]^{1/2},$$

where $0.71 = (1 - 0.29)$ is one minus the squared correlation and $1.32 = n \sum_{i=1}^n w_i^2$. The situation for housekeeping is more extreme. The correlation is not large, and, apparently, the large deviations from the regression line are associated with large weights. The estimated variance for the regression estimator is about twice the estimated variance of the unweighted estimator.

Table 2 also contains the estimated differences between the unweighted and weighted estimators. The difference between the unweighted and the weighted estimated total is

$$\sum_{t=1}^n Nn^{-1}y_t - \sum_{t=1}^n w_t y_t = \sum_{t=1}^n (n^{-1}N - w_t)y_t.$$

The difference between the estimated means is the difference between the totals divided by the population size. To compute the variance of the difference between the means, we note that the hypothesis of a zero difference is equivalent to the hypothesis that the correlation between w_t and y_t is zero. Therefore, we computed the unweighted regression of y_t on w_t and computed the variance of the regression coefficient under the design using PC CARP. The standard errors for the difference in Table 2 are such that the “ t -statistic” for the hypothesis of zero difference is equal to the “ t -statistic” for the coefficient of w_t in the regression of y_t on w_t .

For all four characteristics, the difference between the weighted and unweighted estimators of the population mean is significant at traditional levels. Thus, under the assumption that the regression estimators are unbiased, there are significant biases in the unweighted estimators.

The bias picture is mixed for the estimates of the subpopulation means. The difference between the two estimators is significant for the three means for the housekeeping subpopulation, which is the population of interest. The difference is nonsignificant for the three means for the nonhousekeeping subpopulation. The sample contains only 222 nonhousekeeping households. Therefore, the variance of the difference between the weighted and unweighted estimates is much larger for the nonhousekeeping households than for the housekeeping households.

The differences between the two estimates of the population means are a function of the differences between the two estimates of the subpopulation means and the two estimates of the fraction of households in the two categories. This explains why the difference for “all” can be larger than both the “housekeeping” and “nonhousekeeping” differences.

The last column of Table 2 contains the ratio of the estimated mean square error of the unweighted estimator to the variance of the regression estimator. The estimated mean square errors for the unweighted estimators were computed as

$$\widehat{MSE}_u = \hat{V} + \max\{0, (\text{Diff})^2 - (\text{s.e. diff})^2\},$$

where \hat{V} is the estimated variance of the unweighted estimate, Diff is the difference between the two estimates from Table 2, and s.e. diff is the standard error of the difference from Table 2. The estimated variance \hat{V} for the unweighted estimator is variance formula (2.8) with constant w_{ijt} ,

and with $x_{ijt} \hat{\beta}$ replaced by $\bar{y}_{t..}$. The second term of the estimated mean square error is the estimated squared bias. Under the assumption that the regression estimator is unbiased, the expected value of $(\text{Diff})^2$ is the squared bias plus the variance of the difference. Hence, under the assumption that the regression estimator is unbiased, the estimated mean square error of the unweighted estimator is a consistent estimator. The estimated mean square errors of the weighted estimators are the variances of the weighted estimators computed as the squares of the standard errors of Table 2.

Of the four characteristics for which the population mean was estimated, the estimated relative efficiency of the regression estimator to the simple mean ranges from 2.5 to 129. The regression estimator for meals away has the smallest estimated efficiency. The variances of the two estimators are similar, but because of the estimated bias, the regression estimate for meals away is estimated to have a mean square error that is about 40% of that of the unweighted estimate. The mean square error of the regression estimate for home food is less than 20% of that of the unweighted estimate, that for meal persons is about 1% of that of the unweighted estimate, and that for housekeeping is about 20% that of the unweighted estimator. In all cases, the squared bias is a very important component of the estimated mean square error.

Because the unweighted subpopulation estimates for the nonhousekeeping households showed little bias, the unweighted estimates are estimated to be somewhat more efficient than the regression estimates. The nonhousekeeping subpopulation is only about 6% of the population and the deviations from the subpopulation mean show little correlation with the control variables. On the other hand, the regression estimates for the housekeeping subpopulation are estimated to be much more efficient than the unweighted estimates. The relative efficiencies for the housekeeping subpopulation are close to those of the total population estimates.

Even after allowing for the fact that the population totals from the Current Population Survey are not known population totals, it is clear that large gains are associated with regression estimation for the population means. Although the regression estimator for the means of the small subpopulation is estimated to be less efficient than the unweighted estimators, the loss in efficiency is small relative to the large gains in efficiency estimated for the other variables.

ACKNOWLEDGEMENTS

This research was partly supported by Research Support Agreement 58-3198-9-032 with the Human Nutrition Information Service, U.S. Department of Agriculture. We thank Phil Kott, Patricia Guenther, and the referees for useful comments.

APPENDIX

WEIGHT GENERATION PROGRAM

In this appendix, we present the regression weight generation procedure of Huang and Fuller (1978). The procedure we describe contains the option of specifying maximum and minimum weights. This option was not part of the original program. For a discussion of related weight generation procedures, see Singh (1993).

Suppose that the population means $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$ of the k auxiliary variables (X_1, X_2, \dots, X_k) are known. Let a sample of n observations be available and let

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}, \quad (\text{A.1})$$

where X_{ij} is the observation on variable j for individual i .

In addition to the array of sample observations and the populations means, two initial factors v_i and $g_i^{(0)}$, $i = 1, 2, \dots, n$, are required to initiate the computations. The v_i are typically inversely proportional to the probabilities of selection. The default values for $g_i^{(0)}$ are $g_i^{(0)} = 1$. For stratified samples or data with unequal variances, the user may choose other values for $g_i^{(0)}$. (See Huang 1978 or Goebel 1976.) The program input includes the sample size n , the population size N , the parameter M , the maximum number of iterations LI, the upper bound of the ratios of weights to the average weight U_B , and the lower bound of the ratios of weights to the average weight L_B . It is required that $0 \leq L_B < 1 < U_B$. In our description, we assume $\sum_{i=1}^n v_i = n$. The program normalizes the v_i so that the sum is n .

The program can be used to construct weights to estimate means or to estimate totals. The weights for totals are the weights for the means multiplied by N . For means, the program attempts to construct weights w_i such that

$$\sum_{i=1}^n w_i(1, X_i) = (1, \bar{X}), \quad (\text{A.2})$$

$$L_B < nw_i < U_B, \quad (\text{A.3})$$

$$(1 - M) \max_{1 \leq i \leq n} w_i v_i \leq (1 + M) \min_{1 \leq i \leq n} w_i v_i, \quad (\text{A.4})$$

for $i = 1, 2, \dots, n$.

The program is iterative, where an iteration consists of computing the generalized least squares weights, where a control factor h_i is applied to each observation. The h_i is a product of v_i and g_i , where g_i for iterations after the

first is a "bell" shaped function of the distance (in a suitable metric) that the observation is from the population mean. At each iteration, the weights satisfy (A.2) but may fail (A.3) or (A.4).

It will not always be possible to construct weights satisfying the specified restrictions in the specified number of iterations. If the sample is such that the restriction cannot be met, the program outputs the weights of the last iteration. In the single x case, when the criterion cannot be satisfied, there will be two weights, one for those greater than the population mean, and one for those less than the population mean.

To describe the algorithm, let

$$Z_{ij} = X_{ij} - \bar{X}_j,$$

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1p} \\ \vdots & \vdots & & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{np} \end{pmatrix},$$

$$V = \text{diag}(v_1, v_2, \dots, v_n),$$

$$J_n = (1, 1, \dots, 1)',$$

$$A^{(0)} = Z' H^{(0)} Z,$$

$$G^{(0)} = \text{diag}(g_1^{(0)}, \dots, g_n^{(0)})$$

and

$$H^{(0)} = V G^{(0)}.$$

The algorithm consists of iterating three steps.

1. The initial calculation is for $\alpha = 0$. At iteration α , the vector of regression weights, denoted by $w^{(\alpha)}$, is

$$\begin{aligned} w^{(\alpha)} &= [1 + n\bar{u}_v^{(\alpha)}]^{-1} V(n^{-1}J_n + u^{(\alpha)}) \\ &= (w_1^{(\alpha)}, \dots, w_n^{(\alpha)})', \end{aligned} \quad (\text{A.5})$$

where

$$u^{(\alpha)} = G^{(\alpha)} Z (A^{(\alpha)})^\dagger (\bar{X} - \bar{X}_v) = (u_1^{(\alpha)}, \dots, u_n^{(\alpha)})',$$

$$\bar{X}_v = \left(\sum_{i=1}^n v_i \right)^{-1} \sum_{i=1}^n v_i X_i,$$

$(A^{(\alpha)})^\dagger$ is a symmetric generalized inverse of $A^{(\alpha)}$,

$$n\bar{u}_v^{(\alpha)} = \max\{J_n' V u^{(\alpha)}, n^{-1} - 1\}, \quad (\text{A.6})$$

and

$$A^{(\alpha)} = Z' H^{(\alpha)} Z.$$

2. The weights of Step 1 are checked to see if they satisfy the criteria.

(a) Is $|nu_i^{(\alpha)}| \leq M$ for all i ?

(b) Is

$$L_B \leq nw_i^{(\alpha)} \leq U_B$$

for all i ?

If either (a) or (b) fails for any i and LI iterations have not been completed, go to Step 3. If (a) and (b) are satisfied, or if LI iterations have been completed, the weights are output.

3. The control factors $h_i^{(\alpha)}$, $i = 1, 2, \dots, n$, are modified. Set

$$H^{(\alpha)} = H^{(\alpha-1)}G^{(\alpha)},$$

where

$$G^{(\alpha)} = \text{diag}(g_1^{(\alpha)}, g_2^{(\alpha)}, \dots, g_n^{(\alpha)}),$$

$$\begin{aligned} g_i^{(\alpha)} &= 1 & 0 \leq d_i^{(\alpha)} < 0.5 \\ &= 1 - 0.8(d_i^{(\alpha)} - 0.5)^2 & 0.5 \leq d_i^{(\alpha)} \leq 1 \\ &= 0.8(d_i^{(\alpha)})^{-1} & d_i^{(\alpha)} > 1, \end{aligned}$$

$$d_i^{(\alpha)} = 1.33[D_i^{(\alpha-1)}]^{-1}n|u_i^{(\alpha-1)}|,$$

$$\begin{aligned} D_i^{(\alpha-1)} &= \min\{M, C_{Li}^{(\alpha-1)}\} & \text{if } w_i^{(\alpha-1)} < v_i \\ &= \min\{M, C_{Bi}^{(\alpha-1)}\} & \text{if } w_i^{(\alpha-1)} \geq v_i, \end{aligned}$$

$$C_{Li}^{(\alpha-1)} = \max\{|v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})L_B - 1|, 0.1M\},$$

$$C_{Bi}^{(\alpha-1)} = \max\{|v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})U_B - 1|, 0.1M\}.$$

Go to Step 1 to compute new regression weights.

The constant 1.33 in the definition of $d_i^{(\alpha)}$ and the constant of 0.8 in the definition of $g_i^{(\alpha)}$ were chosen to speed convergence. The control factors $g_i^{(\alpha)}$ are chosen to downweight observations on the basis of a distance from the population mean.

The definition of $w^{(\alpha)}$ in (A.5) is an alternative way of writing the vector of generalized least squares weights of (2.4) when $\pi_i^{-1} = h_i^{(\alpha)}$.

REFERENCES

- AKKERBOOM, J.C., SIKKEL, D., and van HERK, H. (1991). Robust weighting of financial survey data. Contributed paper presented at meeting of the International Statistical Institute, Cairo, Egypt.
- ALEXANDER, C.H. (1987). A model based justification for survey weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.
- BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- BETHLEHEM, J.G., and KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā, Series C*, 97-114.
- BUREAU OF THE CENSUS (1987). Current Population Survey, March 1987: Technical Documentation. Washington, D.C.
- COCHRAN, W.G. (1942). Sampling theory when the units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley.
- COPELAND, K.R., PEITZMEIER, F.K., and HOY, C.E. (1987). An alternative method of controlling current population survey estimates of population counts. *Survey Methodology*, 13, 173-182.
- COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- COX, L.H., and ERNST, L.R. (1982). Controlled rounding. *INFOR*, 20, 423-432.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DARROCH, J.N., and RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43, 1470-1480.
- EL-BADRY, M.A., and STEPHAN, F.F. (1985). On adjusting sample tabulations to census counts. *Journal of the American Statistical Association*, 50, 738-762.
- FAGAN, J.T., GREENBERG, B.V., and HEMMIG, B. (1988). Controlled rounding of three dimensional tables. Statistical Research Division Report Census/SRD/RR-88/02. U.S. Bureau of the Census, Washington, D.C.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C*, 37, 117-132.
- FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames Iowa.

- GOEBEL, J.J. (1976). Application of an iterative regression technique to a national potential cropland survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 350-353.
- HAMPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the Statistical Computing Section, American Statistical Association*, 59-64.
- HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.
- HUANG, E.T. (1978). Nonnegative regression estimation for sample survey data. Unpublished Ph. D. thesis. Iowa State University, Ames, Iowa.
- HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Social Statistics Section, American Statistical Association 1978*. 300-303.
- HULLIGER, B. (1993). Robustification of the Horvitz-Thompson estimator. Contributed paper 49th Session of the International Statistical Institute. Book 1, 583-584.
- HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.
- IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 169-188.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Experiment Station Research Bulletin, 304.
- KRASKER, W.A. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, 48, 1333-1346.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- LUERY, D. (1986). Weighting survey data under linear constraints on the weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 325-330.
- MALLOWS, C.L. (1983). Discussion of Huber: Mimimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78, 77.
- MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.
- OH, H.L., and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.
- RAO, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. Paper presented at the Workshop on Users of Auxiliary Information in Surveys, Örebro, Sweden, October, 1992.
- ROYALL, R.M., and CUMBERLAND, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SAS INSTITUTE INC. (1989). SAS/STAT User's Guide, Version 6 Fourth Edition, Volume 1. Cary, NC: SAS Institute Inc.
- SINGH, A.C. (1993). On weight adjustment in survey sampling. Unpublished manuscript. Statistics Canada, Ottawa, Canada.
- STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.
- WATSON, D.J. (1937). The estimation of leaf areas. *Journal of Agricultural Science*, 27, 474-483.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.
- ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

Estimating the Rate of Rural Homelessness: A Study of Nonurban Ohio

ELIZABETH A. STASNY, BEVERLY G. TOOMEY and RICHARD J. FIRST¹

ABSTRACT

Recently, much effort has been directed towards counting and characterizing the homeless. Most of this work, however, has focused on homeless persons in urban areas. In this paper, we describe efforts to estimate the rate of homelessness in nonurban counties in Ohio. The methods for locating homeless persons and even the definition of homelessness are different in rural areas where there are fewer institutions for sheltering and feeding the homeless. There may also be a problem with using standard survey sampling estimators, which typically require large population sizes, large sample sizes, and small sampling fractions. We describe a survey of homeless persons in nonurban Ohio and present a simulation study to assess the usefulness of standard estimators for a population proportion from a stratified cluster sample.

KEY WORDS: Biased estimator; Regression estimator; Small sample size; Stratified cluster sample; Simulation.

1. INTRODUCTION

When we think of the homeless, we often think of "street people" and "bag ladies". We picture people sleeping on park benches, on heating grates, and in homeless shelters. These stereotypes of the homeless originated in large cities, however, and do not necessarily provide an accurate picture of homeless persons in rural areas.

Many of the studies of homeless persons have been carried out in larger cities. For example, the 1987 Urban Institute Study counted homeless persons in 20 major cities in the U.S. Another major study by Rossi was carried out in Chicago. (See Burt and Taeuber (1991) for an overview of survey methods for these and other studies that counted homeless populations.)

During the 1990 United States Population Census, a special attempt was made to include homeless persons in the population count through the S-Night (Shelter and Street Night) count. For this effort, a special national list of shelters and locations in which homeless persons sleep was compiled. The highest elected official of over 39,000 rural and urban local governments was asked to provide a list of shelters, street locations, and open public locations where the homeless stay at night. The homeless were counted by Census enumerators during a single night, March 20. Note that the main goal of S-Night was to include homeless persons in the Census count; relatively little information on characteristics of the homeless is available in the Census data. Details on the S-Night procedures are provided by Taeuber and Siegel (1990).

In contrast to surveys of homeless persons in urban areas and to the Census S-Night, the goal of the survey described here was to locate and count the nonurban

homeless wherever they might be, and to collect information to describe these homeless persons. In Section 2 of this paper, we describe the design of the 1990 survey of rural homeless persons in Ohio. We present our definition of rural homelessness and we describe the methods used to locate and survey the homeless. In Section 3 we present our estimates of the rates of rural homelessness obtained using the standard estimator of a proportion from a stratified cluster sample. Since these estimates are likely to be biased, we also present the results of a simulation study conducted to assess the likely size of the bias. In Section 4 we consider a regression estimator for the rate of homelessness and compare the regression estimator to the standard estimator of Section 3. In Section 5 we present our conclusions.

2. THE SURVEY

There are 88 counties in Ohio. Of these, 13 are urban counties with large cities and 75 are defined as rural or nonurban. These 75 counties of interest include counties that are completely rural, counties that are not adjacent to urban counties and that have moderately populated county seats, and suburban counties that border counties with large metropolitan areas.

The design used in this 1990 survey was selected to facilitate comparisons with a 1984 study of Ohio rural homeless persons (Roth *et al.* 1985). In the earlier study, Ohio's counties were divided into five regions, northeast, northwest, central, southeast, and southwest, and a stratified random sample of 16 rural counties was selected. The 21 counties selected for the 1990 study included the

¹ Elizabeth A. Stasny, Department of Statistics; Beverly G. Toomey and Richard J. First, College of Social Work, The Ohio State University, Columbus, Ohio 43210.



Note: Shaded counties are urban counties that were excluded from the study. An "S" indicates a county in the sample. The heavy boundaries divide the state into the five geographical strata: northeast, northwest, central, southeast, and southwest.

Figure 1. County Map of Ohio

16 counties from the original study and one additional county selected at random from within each region. (We should note that analysis of data from the present study suggests that stratification of Ohio into the five regions is not useful for improving the estimate of rural homelessness.) A map of Ohio showing the five regions, the urban counties, and the sampled counties is provided in Figure 1.

The following is a brief description of the 1990 survey methodology. More detailed descriptions are given by First *et al.* (1994) and Toomey *et al.* (1993).

2.1 Survey Personnel

A census of all homeless persons within the 21 sampled counties was attempted. Because there are not typically homeless shelters or other gathering places for the homeless in nonurban areas, the survey was conducted over a six-month period and made use of a network of advisors to locate the homeless. The survey period began with the first full week of February 1990. Homeless persons were identified and located by a referral network within each sampled county. Each network was supervised by a local county coordinator. The principal investigators supervised

the county coordinators and the central office staff. They monitored the data collection, through bi-weekly phone calls and field visits, to assure uniformity and to control quality.

Advisors and interviewers, selected for their knowledge of the counties in which they worked, identified people who met the criteria for homelessness. Advisors included church leaders, hospital staff, civic club leaders, elected community officials, informal community leaders, bartenders, hotel clerks, laundromat attendants, and professional service providers such as health department staff, librarians, agricultural extension agents, postal workers, ministers, park rangers, neighborhood action groups, human service case workers, mental health workers, and law enforcement officers. One hundred interviewers conducted the interviews with the homeless. Interviewers attended a four-hour training session and were provided with a training manual of field guidelines. Interviews took place in offices, diners, motel rooms, cars, state parks, barns, laundromats, bars, and under railroad trestles. Interviewers were trained to know about available community resources and to make referrals for respondents who wanted services. In addition, interviewers had access to funds to offer a meal or minor assistance if necessary (less than \$600 was spent on such purchases). Assistance provided through interviewers was limited so that people would not have an incentive to falsely identify themselves as homeless.

2.2 Definition of Rural Homeless

Screening questions were used to identify homeless persons. The definition of homelessness used in this study was necessarily somewhat different from the definition used for studies in urban areas. In rural areas there are fewer public shelters and housing alternatives specifically for the homeless. Respondents were counted as homeless if they did not have a permanent residence they considered home and if, on the previous night, they had slept in (1) limited or no shelter, (2) shelters or missions that serve homeless persons, (3) cheap hotels or motels when the actual stay or intent to stay was 45 days or less, or (4) other unique situations when the actual stay or intent to stay was 45 days or less. Included in the fourth category were people who stayed in sheds, barns, old buses, and old trailers without water or power, provided the person did not own the property and was not paying rent to stay there. Also included as homeless were people who were temporarily staying with friends or relatives, had not been staying in that household more than 45 days, were not a part of the household, and were planning on moving out in 45 days or less. Persons who were staying in battered women's shelters, hospitals, prisons, migrant workers camps, *etc.* were not counted as homeless unless they were leaving the facility and had nowhere to go.

Our definition of homelessness may be contrasted with that used in studies of homeless persons in urban areas. The common criteria of the definition of homelessness for such studies is based on the Stewart B. McKinney Homeless Assistance Act (1987). The Act defines a homeless person as "an individual who lacks a fixed, regular, and adequate nighttime residence and an individual who has a primary nighttime residence that is (a) a supervised publicly or privately operated shelter designed to provide temporary living accommodations (including welfare hotels, congregate shelters, and transitional housing for the mentally ill); (b) an institution that provides a temporary residence for individuals intended to be institutionalized; or (c) a public or private place not designed for, or ordinarily used as, a regular sleeping accommodation for human beings." From this definition comes the notion of "literally homeless" as suggested by Rossi *et al.* (1987). These standard definitions do not include, for example, those homeless persons who double up with family or friends. We did include such persons in our count of the rural homeless. Our analysis indicates that about a third of the persons counted in our census would not be counted under the urban definition of homelessness. It is not known how much counting those doubling up would increase estimates in urban areas.

2.3 The Interview Period

The use of a six-month survey period for counting the rural homeless is different from the typical one-day survey period used most often in surveys conducted in urban areas. In a review of seven studies of the homeless, Burt and Taeuber (1991) report that these studies used single nights, or one or two weeks as the interview period at a single location. Most of these studies relied on locating the homeless in shelters, soup kitchens, abandoned buildings, or similar locations. Since the homeless in rural areas are less likely to have shelters or soup kitchens available to them, they are harder to find and a longer survey period is recommended.

To facilitate comparisons with single-day or single-week surveys, homeless persons found in this study were asked how long they had been homeless. Using this information we were able to determine the number of persons in the sampled counties who were homeless during the first week of the survey, the first full week of February 1990.

In Section 3 we present estimates of the homeless rate for both the six-month period and the single week. The six-month rate includes anyone who met the definition of homelessness at any time during the six-month interview period. The one-week rate includes those interviewed throughout the six months who reported being homeless during the first full week of February.

To avoid duplication of respondents over the six-month period, each subject was assigned a unique identification number which included the subject's birth date, gender,

and first three letters of the last name. Only a single duplicate interview was found in the data base; it was removed from the data base. (We do not have information on duplicates found in the field.) Because of this control for duplicate counting, we feel that any bias in our data collection procedures would be in the direction of an undercount of the rural homeless.

During the six-month interviewing period, 1,100 adults and 480 accompanying children were identified as homeless in the 21 sampled counties.

2.4 The Survey Questionnaire

If the responses to the screening questions indicated that a person was homeless, that subject was asked to respond to a questionnaire designed to obtain information about the person and his or her life experiences. Of the 1,100 adults identified as homeless, 919 completed the full interview. Although the focus of this paper is on estimating the number of rural homeless, we will describe briefly the questionnaire used to collect information to characterize the homeless.

The full questionnaire contained three sections. The first included questions on demographics and life experiences (for example, reasons for being homeless, use of mental health and other human services, employment history, drug and alcohol usage, family structure, and general well-being). The second section contained ten scales (including, for example, depression-anxiety, disorientation-memory impairment, and retardation-lack of emotion) from the Psychiatric Status Schedule developed by Spitzer, *et al.* (1970). The final section was an interview post-mortem which was completed by the interviewer and included information on where the interview occurred, respondent characteristics (for example, gender and unusual behaviors), and an assessment of the accuracy of the respondent's answers. The findings from this portion of the study are summarized by First *et al.* (1994).

3. THE ESTIMATES OF RATE OF HOMELESSNESS

3.1 The Estimator

The regional estimate of the rate of rural homelessness was obtained using the standard estimator for a proportion from a stratified cluster sample with unequal cluster sizes. In this case, the cluster is the county, the cluster size is the population within the county, and the strata size is the population within a region. The estimator is as follows:

For the i -th region, the estimated rate of homelessness is r_i where

$$r_i = \frac{\text{number of homeless in sampled rural counties in the } i\text{-th region}}{\text{total population in sampled rural counties in } i\text{-th region}}$$

Then the estimated homeless rate for the state is

$$r_{\text{state}} = \frac{\sum_i [r_i \times \text{rural county population in } i\text{-th region}],}{\text{total rural county population in Ohio}}$$

where the summation is over the five geographical regions shown in Figure 1. The population totals for the 75 non-urban counties were obtained from 1990 Census data.

The estimated one-week and six-month rates of homelessness, given as number of homeless persons per 10,000 population, are shown in Table 1.

Because the above estimator involves the ratio of two random variables, the number of homeless and the population size for sampled clusters, the estimator is biased (see, for example, Cochran 1977). The bias decreases as sample size (number of counties sampled in this case) increases. Since our sample size is small, we recognize that our estimates are likely to be biased. On the other hand, our sampling fraction is large because the number of rural counties is small. Hence, we wish to assess the likely amount of bias in our estimates. (Note that the small sample sizes could also make the standard errors given in Table 1 inaccurate.)

Table 1
Estimated Rates of Homelessness per 10,000
in Rural Ohio

Area	One-Week Rate (February 4 – February 10, 1990)	Six-Month Rate (February – July 1990)
State	5.68 (0.99)	14.00 (2.09)
Northeast	3.44 (0.79)	12.00 (2.19)
Northwest	5.21 (3.51)	12.77 (5.18)
Central	5.85 (1.86)	12.11 (3.05)
Southeast	6.89 (1.93)	15.90 (5.91)
Southwest	7.25 (2.44)	16.78 (5.32)

Note: Standard errors are given in parentheses after each estimate.

3.2 The Simulation Study

We conducted a simulation study to help us assess the likely amount of bias in our estimates. We first generated five "populations" each with counts of the homeless for all 75 nonurban counties in Ohio. For all five simulated populations, the observed numbers of homeless persons for the 21 sampled counties were used as the counts in those counties. Counts for the remaining 54 counties were generated randomly as described below. Note that the simulated counts represent the six-month counts of the homeless.

The first simulated population was created by generating the natural log of the rate of homelessness from a single normal distribution. The log of the rate was used because the observed rates for the 21 sampled counties have a highly skewed histogram but the histogram for the log of the rates is approximately mound shaped. The mean of the observed log rates is 2.465 with a standard deviation of 0.7154. Thus, the generated log rates of homelessness were randomly sampled (using the statistical package S) from a normal distribution with this mean and standard deviation. After the log rates were generated for the 54 nonsampled counties, they were used along with the population counts from the 1990 Census for each county to obtain the simulated numbers of homeless persons for those counties.

The second simulated population was created in a manner similar to the first except that separate normal distributions were used within each of the five geographic regions of Ohio. The means and standard deviations of the log rates of homelessness for the sampled counties within each region were used as the parameters of the normal distributions from which the simulated values were generated. Again the simulated log rates were used to obtain the numbers of homeless persons for the 54 nonsampled rural counties.

The third simulated population was generated using the regression of rate of homelessness per 10,000 on the percent elderly in each sampled county. (This choice of predictor variable is based on the selection of a regression estimator as described in Section 4.) The fitted regression model is

$$\widehat{\text{rate}} = -9.02 + 2.32\% \text{elderly},$$

with $R^2 = 0.197$, $\sqrt{\text{MSE}} = 9.03$, and $p\text{-value} = 0.044$ for the overall F-test for the regression line. The simulated population was created by estimating the rate of homelessness in each nonsampled county from the percent elderly in the county and then adding a random normal error term. Because a plot of the residuals from the regression line suggested that the variance in the residuals is larger for counties with higher percentages of elderly, the random error terms were generated from two different normal distributions depending on whether the percent elderly in the county was more or less than 10%. The standard deviations used for the two normal distributions were the standard deviations in the residuals for the counties with 10% or more elderly and with less than 10% elderly.

The fourth simulated population was generated using the regression of rate of homelessness per 10,000 on the percent elderly in each sampled county and on the indicators of the region of the state to which the county belongs. Using I_{NE} , I_{NW} , I_C , and I_{SE} to represent indicator variables for the northeast, northwest, central, and southeast regions respectively, the fitted regression model is

$$\begin{aligned} \widehat{\text{rate}} = & -10.40 + 3.23\% \text{elderly} \\ & - 6.47 I_{NE} - 8.55 I_{NW} - 8.64 I_C - 14.25 I_{SE}, \end{aligned}$$

with $R^2 = .407$ ($R^2\text{-adjusted} = .210$), $\sqrt{\text{MSE}} = 8.73$, and $p\text{-value} = 0.127$ for the overall F-test for the regression line. The simulated population was created by estimating the rate of homelessness in each nonsampled county from the regression equation and then adding a random normal error term. A residual plot again suggested that the variance in the residuals is larger for counties with higher percentages of elderly. Thus the random error terms were generated from two different normal distributions depending on whether the percent elderly in the county was more or less than 10%. Again, the standard deviations for the two normal distributions were the standard deviations of the appropriate subsets of residuals.

The fifth simulated population was generated to be somewhat different from the other populations. It was generated using a regression model to predict number of homeless directly from the population size within each county. The fitted regression model is

$$\widehat{\text{homeless}} = 13.23 + 0.001154 \text{population},$$

with $R^2 = 0.386$, $\sqrt{\text{MSE}} = 54.29$, and $p\text{-value} = 0.003$ for the overall F-test for the regression line. The simulated population was created by estimating the number of homeless persons in each nonsampled county from the fitted regression equation and then adding a random normal error term. Because a plot of the residuals suggested that the variance in the residuals is larger for counties with larger populations, the random error terms were generated from two different normal distributions depending on whether the county population was more or less than 30,000. The standard deviations for the two normal distributions were the standard deviations of the appropriate subsets of residuals.

After the five populations had been generated, they were each used to assess the amount of bias in the estimates of the rate of rural homelessness. Since we had created the entire "population", we could compute the "true" rate of homelessness within the entire state and the five geographical regions for each of the five populations.

In the simulation, samples of 21 rural counties were selected using the stratified sampling scheme that was used for the actual study. That is, four counties were sampled at random without replacement from each of the northeast, northwest, central, and southwest regions; five were sampled from the southeast region. The estimated rates of homelessness were computed for the five regions and for the state using the formulas given in Section 3.1. These estimates were compared to the population rates of homelessness for the simulated population to determine the bias in the estimate. This process of selecting a sample,

computing estimates, and determining the bias was repeated 1 million times with replacement for each simulated population. (The number of possible samples is more than 7.15×10^{15} .) The same stream of random numbers was used to select the samples for each of the five populations. The results of the simulation are presented in Table 2.

Table 2

Bias in the Estimate of the Homeless Rate per 10,000
for Five Simulated Populations
(Based on 1,000,000 simulated samples)

	Population				
	1	2	3	4	5
STATE	0.0406 (2.056)	0.1308 (1.759)	0.2618 (2.144)	0.2433 (1.807)	0.2547 (1.605)
REGION					
NE	-0.0406 (3.333)	-0.0379 (2.923)	0.1538 (3.748)	0.0317 (4.034)	0.0993 (1.937)
NW	-0.0578 (3.591)	-0.2948 (3.194)	0.0529 (3.474)	0.0254 (3.460)	0.3234 (4.249)
C	-0.2442 (3.122)	0.2700 (3.762)	0.3974 (3.426)	0.1362 (2.260)	0.1901 (2.869)
SE	-0.1034 (6.512)	-0.0279 (4.298)	-0.1132 (6.600)	-0.1798 (3.892)	0.0427 (3.973)
SW	0.6184 (4.215)	0.8093 (4.990)	0.9196 (4.610)	1.277 (5.173)	0.6716 (4.274)

Note: The standard deviation of the simulated sampling distribution of the estimator is given in parentheses below each value.

From Table 2 we see that the size of the bias in the overall state estimate of homelessness is about 1/100th of the size of the estimate itself. (Recall that the actual estimated six-month rate of homelessness for the state is about 14 per 10,000 population. The simulated populations have state rates between about 13 and 15 per 10,000.) At the regional level, the size of the bias is also about 1/100th of the size of the regional estimates even though the regional estimates are based on much smaller sample sizes. These results suggest that the size of the bias in our actual estimate is likely to be relatively small.

As would be expected from the small number of counties in the sample, the variance of the sampling distribution of the estimator is fairly large. The standard deviation in the estimates from the simulation study was about 10 times the size of the bias. (The standard deviations of the 1,000,000 estimates in each of the five simulations are of the same order of magnitude as the standard error of the actual estimate shown in Table 1.) This result suggests that the bias in the actual estimate is likely to be rather unimportant when compared to the standard error of the estimate.

Finally, we assessed the shape of the sampling distribution of our estimator by looking at histograms of the 1,000,000 estimates from each of our five simulation studies. The histograms appeared symmetric, mound shaped, and remarkably like histograms of normal data. Thus, confidence intervals based on the normal approximation are likely to be fairly accurate.

4. A REGRESSION ESTIMATOR

There is a great deal of information available, for example from the Bureau of the Census, on the economic conditions in a county. We hoped to be able to use some of this information to improve our estimate for the rate of homelessness by using a regression estimator. To this end, we searched for a regression model relating either the number of homeless persons in a county or the rate of homelessness with a variety of predictor variables which we thought might be useful in explaining homelessness. These possible predictor variables included county population, percentage change in population from 1980 to 1990, unemployment rate, percent elderly, public welfare expenditures, average weekly earnings, percent of rental property, median rent, poverty rate, percent female head of household, percentage of land in farming, average value of farms, average income per farm, ratio of manufacturing to farm jobs, indicator of Beale scores – a classification system for degree of ruralness (see Thomas 1977), and regional indicators.

None of these possible predictors individually or in combination provided a good predictor of the number of homeless persons or rate of homelessness. The best single predictor was percent elderly, the model which was used in generating the third simulated population described in Section 3.2, but it explained less than 20% of the variability in the rate of homelessness. No other variable was useful in addition to percent elderly and we could not find another reasonable regression model. Thus we used percent elderly in a regression estimator for the state rate of rural homelessness. Note that percent elderly is a plausible predictor of the rate of homelessness because poor economic conditions in a rural county appear to result in out-migration of the young; the elderly remain behind making up a greater proportion of the population. Therefore, it is possible that the percentage of elderly in a county is a proxy for poor economic conditions and out-migration. We cannot, however, rule out the possibility that percent elderly appears to be related to rate of homelessness in our data due to chance. We also realize that unavoidable errors in the county-based data collection procedures, such as interviewer effect, amount of services available, and geographic factors, may contribute to the lack of association between rate of homelessness and theoretically relevant variables.

We used the combined regression estimator (see, for example, Cochran 1977) to obtain the state estimate of 14.85 rural homeless per 10,000 with a standard error of 1.64. This compares with the original estimate of 14.00 with a standard error of 2.09 as shown in Table 1. Because the regression estimator is also biased with the bias decreasing for larger sample sizes, we again used a simulation study to assess the bias in this regression estimator.

The simulation study for the regression estimator was carried out using the third and fourth simulated populations described in Section 3.2 because those populations were generated using a regression model involving percent elderly. The simulation again computed the bias in the estimate for 1 million stratified cluster samples chosen with replacement from each population. The same stream of random numbers was used to generate the samples in both cases. A summary of the results of the simulation study for both the original estimator and the regression estimator is given in Table 3.

Table 3

Comparison of Estimators of State Homeless Rate per 10,000
(Summary for 1,000,000 repetitions from
two simulated populations)

	Original Estimator		Regression Estimator	
	Population		Population	
	3	4	3	4
Average Bias	0.2618	0.2433	1.7115	0.8360
Standard Deviation	2.144	1.807	1.820	1.246
MSE	4.664	3.325	6.242	2.250

Note that the average bias is larger for the regression estimator than for the standard estimator for a rate from a stratified cluster sample. The standard deviation of the sampling distribution for the regression estimator, however, appears to be slightly smaller than that of the original estimator for each of the two simulated populations. The mean squared errors for the regression estimator fell above and below those of the original estimator. Thus, the choice of which estimator to use was unclear from the summary information in Table 3.

Because the regression estimator does not provide a clear improvement over the original estimator, the bias on average appears to be larger for the regression estimator, and the percent elderly variable may have been selected out of the many variables we tried due to chance, we chose to use the standard estimator of Section 3 for estimating the rate of rural homelessness.

5. CONCLUSIONS

The most often quoted national figures on homelessness were published by Burt and Cohen (1989) who estimated rates of homelessness in urban areas at 37.4 per 10,000 population in cities of more than 100,000 and 9 per 10,000 outside of SMAs. This current study of homeless persons in nonurban Ohio gives a six-month rate of about 14 homeless per 10,000 population and a one-week rate of 5.68 per 10,000 population.

The results of our simulation study suggest that the bias in the usual estimate of a rate based on our small cluster sample is not likely to be important, particularly in comparison to the size of the standard error of the estimate. The bias in the estimates for the five geographic regions in Ohio was found to be of a similar, relatively small size. The simulation study suggests that statistical biases and errors are not likely to discredit the substantive results of the survey of rural homeless.

Our regression analysis of the numbers of homeless persons from sampled counties suggests that it is difficult to explain the numbers of homeless persons in nonurban counties using economic and demographic variables that might be thought to be related to homelessness. It may be that each county is so different from the others, because of its location relative to population centers and related economic characteristics, that it is impossible to find a suitable stratification of the nonurban counties within Ohio. The use of a geographically stratified sample in Ohio did not appear to reduce the variance of the estimate and no other stratification variable was suggested by our regression analysis. This may be the case for other states as well, although stratification by some variable may be possible over, say, the entire United States.

ACKNOWLEDGEMENTS

This research was supported in part by Grant #R01MH46111 from the National Institute of Mental Health and by the Ohio Departments of Health and Mental Health. The authors wish to recognize the excellent work of Ms. H. C. Tsai who did most of the computing for this research. The authors thank two referees for their useful comments on an earlier version of this paper.

REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd edition). New York: Wiley.
- BURT, M.R., and COHEN, B. (1989). *America's Homeless: Numbers, Characteristics and Programs that Serve Them*. Urban Institute Report 89-3. Washington, DC: Urban Institute Press.

- BURT, M.R., and TAEUBER, C. M. (1991). Overview of seven studies that counted or estimated homeless populations in *Enumerating Homeless Persons: Methods and Data Needs*. Proceedings of Census Bureau/ Interagency Council/ Department of Housing and Urban Development Conference, November 29-30, 1990, edited by C.M. Taeuber, Washington, D.C.: US Department of Commerce, 30-76.
- FIRST, R.J., TOOMEY, B.G., RIFE, J.C., and STASNY, E.A. (1994). *Outside of the City: A Statewide Study of Homelessness in Nonurban/Rural Areas*. Final report for NIMH Grant #R01MH46111. Columbus, OH: College of Social Work, The Ohio State University.
- ROSSI, P.H., WRIGHT, J.D., FISCHER G.A., and WILLIS, G. (1987). The urban homeless: estimating composition and size. *Science*, 235, 1336-1341.
- ROTH, D., BEAN, J., LUST, N., and SAVEANU, T. (1985). *Homelessness in Ohio: A Study of People in Need*. Ohio Department of Mental Health, Columbus, Ohio.
- SPITZER, R., ENDICOTT, J., and COHEN, J. (1970). The psychiatric status schedule: A technique for evaluation psychopathology and impairment in role functioning. *Archives of General Psychiatry*, 23, 41-55.
- TAEUBER, C.M., and SIEGEL, P.M. (1990). Counting the Nation's Homeless Population in the 1990 Census. Paper presented at the Conference on Enumerating Homeless Persons: Methods and Data Needs, Washington, D.C., November 29, 1990.
- THOMAS, D.W. (1977). Beale Code Revisions Based on Those Devised by Fuguitt and Beale. The Rural Turnaround in Ohio: 1970-1975. E.S.S. #560. Columbus, OH: Department of Agricultural Economics and Rural Sociology, The Ohio State University.
- TOOMEY, B.G., FIRST, R.J., GREENLEE, R., and CUMMINS, L. (1993). Counting the rural homeless population: Methodological dilemmas. *Social Work Research and Abstracts*, 24, 23-27.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; l, I).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

L'analyse de régression que nous avons effectuée afin d'estimer le nombre de sans-abri dans les comtés échantonnés donne à penser qu'il est difficile d'expliquer le nombre de sans-abri dans des comtés non urbains à l'aide de variables économiques et démographiques dont on pourrait croire qu'elles ont un rapport avec la clochardise. La difficulté vient peut-être du fait que les comtés sont tellement différents les uns des autres, à cause de leur situation géographique par rapport aux agglomérations et des caractéristiques économiques propres à cette situation, qu'il est impossible de trouver une formule convenable pour la stratification des comtés non urbains de l'Ohio. L'utilisation d'un échantillon stratifié selon un critère géographique ne semble pas réduire la variance de l'estimation et notre analyse de régression n'indique pas d'autre variable de stratification possible. On pourrait bien observer la même chose pour d'autres États, quoiqu'il soit possible d'effectuer une stratification selon une variable quelconque pour l'ensemble des États-Unis par exemple.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à une subvention du National Institute of Mental Health (n° R01MH46111) et grâce aux départements de la santé et de la santé mentale de l'Ohio. Les auteurs tiennent à souligner l'excellent travail de Mme H.C. Tsai, qui a effectué la plupart des calculs nécessaires à cette étude. Ils remercient aussi deux arbitres pour les commentaires utiles qu'ils ont formulés sur une version antérieure de cet article.

BIBLIOGRAPHIE

COCHRAN, W.G. (1977). *Sampling Techniques*, (3e édition). New York: Wiley.

BURT, M.R., et COHEN, B. (1989). *America's Homeless: Numbers, Characteristics and Programs that Serve Them*. Urban Institute Report 89-3. Washington, DC: Urban Institute Press.

BURT, M.R., et TAEBER, C. M. (1991). Overview of seven studies that counted or estimated homeless populations dans *Enumerating Homeless Persons: Methods and Data Needs*. Proceedings of Census Bureau/ Interagency Council/ Department of Housing and Urban Development Conference, November 29-30, 1990, édité par C.M. Tauber, Washington, D.C.: US Department of Commerce, 30-76.

FIRST, R.J., TOOMEY, B.G., RIFE, J.C., et STASNY, E.A. (1994). Outside of the City: A Statewide Study of Homelessness in Nonurban/Rural Areas. Final report for NIMH Grant #R01MH46111. Columbus, OH: College of Social Work, The Ohio State University.

ROSSI, P.H., WRIGHT, J.D., FISCHER, G.A., et WILLIS, G. (1987). The urban homeless: estimating composition and size. *Science*, 235, 1336-1341.

ROTH, D., BEAN, J., LUST, N., et SAVEANU, T. (1985). *Homelessness in Ohio: A Study of People in Need*. Ohio Department of Mental Health, Columbus, Ohio.

SPITZER, R., ENDICOTT, J., et COHEN, J. (1970). The psychiatric status schedule: A technique for evaluation psychopathology and impairment in role functioning. *Archives of General Psychiatry*, 23, 41-55.

TAEBER, C.M., et SIEGEL, P.M. (1990). Counting the Nation's Homeless Population in the 1990 Census. Document présenté au Conference on Enumerating Homeless Persons: Methods and Data Needs, Washington, D.C., November 29, 1990.

THOMAS, D.W. (1977). Beale Code Revisions Based on Those Devised by Fugitt and Beale. The Rural Turnaround in Ohio: 1970-1975. E.S.S. #560. Columbus, OH: Department of Agricultural Economics and Rural Sociology, The Ohio State University.

TOOMEY, B.G., FIRST, R.J., GREENLEE, R., et CUMMINS, L. (1993). Counting the rural homeless population: Methodological dilemmas. *Social Work Research and Abstracts*, 24, 23-27.

Aucune de ces variables, ni aucune combinaison de celles-ci, ne s'est révélée un bon prédicteur du nombre de sans-abri ou du taux de clochardise. La seule variable qui soit digne de mention est la proportion des personnes âgées dans la population, qui a d'ailleurs servi à générer la troisième population simulée décrite dans la section 3.2, mais cette variable explique moins de 20% de la variation du taux de clochardise. Aucune autre variable ne pouvait expliquer convenablement la clochardise et nous n'avons pu trouver d'autre modèle de régression qui soit acceptable. C'est pourquoi nous sommes servis de la proportion des personnes âgées dans l'estimateur par régression du taux de sans-abri en région rurale pour l'Etat. Notons que cette variable est un prédicteur plausible du taux de sans-abri car l'existence de conditions économiques difficiles dans un comté rural semble être un motif de migration pour les jeunes; les personnes âgées, elles, ne bougent pas et leur poids démographique dans le comté augmente graduellement. C'est pourquoi la proportion de personnes âgées dans la population d'un comté peut être un indicateur d'une conjoncture difficile et d'une émigration interne. Cependant, nous ne pouvons exclure l'hypothèse que le lien observé dans nos données entre cette variable et le taux de sans-abri soit uniquement l'effet du hasard. Nous reconnaissons aussi que les erreurs qui surviennent inmanquablement dans la collecte des données sur le terrain (par ex., effet de l'intervieweur, quantité de services offerts, facteurs géographiques) peuvent atténuer le lien qui devrait exister théoriquement entre le taux de sans-abri et certaines variables.

L'estimateur par régression combiné (voir, par exemple, Cochran (1977)) a donné un taux de sans-abri en région rurale (pour l'Etat) de 14.85 par tranche de 10,000 habitants, avec une erreur type de 1.64. Cette valeur se rapproche de l'estimation qui figure dans le tableau 1 pour la période de six mois, c'est-à-dire 14.00, avec une erreur type de 2.09. Comme l'estimateur par régression est lui aussi un estimateur biaisé, dont le biais diminue à mesure que la taille de l'échantillon augmente, nous avons effectué de nouveau une étude de simulation pour évaluer ce biais.

Comparaison d'estimateurs du taux de sans-abri pour l'Etat (par tranche de 10,000 habitants) (données basées sur 1,000,000 d'échantillons tirés à répétition dans deux populations simulées)					
Estimateur de la section 3.1	Population		Population	Estimateur par régression	
	3	4		3	4
Biais moyen	0.2618	0.2433	1.7115	0.8360	
Ecart type	2.144	1.807	1.820	1.246	
BQM	4.664	3.325	6.242	2.250	

5. CONCLUSIONS

Nous nous sommes servis des troisième et quatrième populations simulées de la section 3.2 pour réaliser cette étude de simulation parce que ces populations avaient été générées à l'aide d'un modèle de régression qui faisait intervenir la proportion des personnes âgées. Comme pour la simulation de la section 3.2, nous avons calculé le biais de l'estimateur pour 1 million d'échantillons en grappes stratifiées tirés avec remise dans chaque population. La même série de nombres aléatoires a servi à la formation des échantillons pour les deux populations. Les résultats de l'étude de simulation pour l'estimateur de la section 3.1 et l'estimateur par régression sont résumés dans le tableau 3.

On remarque que l'estimateur par régression a un biais moyen plus élevé que l'estimateur de proportion basé sur un échantillon en grappes stratifié. En revanche, l'écart type de la distribution d'échantillonnage de l'estimateur par régression semble légèrement plus petit que celui de la distribution d'échantillonnage de l'autre estimateur pour chacune des deux populations simulées. Pour ce qui est de l'erreur quadratique moyenne (EQM), l'estimateur par régression est supérieur à l'estimateur de proportion dans un cas mais moins efficace dans l'autre. Par conséquent, les résultats du tableau 3 ne nous permettent pas de déterminer clairement quel estimateur conviendrait le mieux.

Puisque l'estimateur par régression ne représente pas une amélioration notable par rapport à l'estimateur de proportion et que son biais semble en moyenne plus élevé, et puisqu'il se peut que le choix de la variable "proportion des personnes âgées" (parmi tant d'autres) ait été uniquement l'effet du hasard, nous avons décidé d'utiliser l'estimateur de la section 3 pour le calcul du taux de sans-abri en région rurale.

Les données nationales sur la clochardise le plus souvent citées se trouvent dans Burt et Cohen (1989); ce rapport estime le taux de sans-abri en région urbaine à 37.4 par tranche de 10,000 habitants dans les villes de plus de 100,000 habitants et à 9 par tranche de 10,000 à l'extérieur des SMA (Standard Metropolitan Areas). L'étude qui a été faite sur les sans-abri dans les régions non urbaines de l'Ohio donne un taux de clochardise d'environ 14 par tranche de 10,000 pour une période de six mois et un taux de 5.68 par tranche de 10,000 pour une période d'une semaine.

Les résultats de notre étude de simulation permettent de croire que le biais de l'estimateur ordinaire de la proportion basé sur un petit échantillon en grappes est probablement négligeable si on le compare surtout à l'erreur type de l'estimation. Le biais des estimations relatives aux cinq régions géographiques de l'Ohio est, lui aussi, relativement faible. D'après les résultats de l'étude de simulation, il n'y a pas lieu de croire que les biais et les erreurs statistiques gênent les principaux résultats de l'enquête sur les sans-abri en région rurale.

On voit d'après le tableau 2 que le biais de l'estimation du taux de sans-abri pour l'État équivalait environ au centième de cette valeur estimée. (Rappelons que le taux estimé de sans-abri pour la période de six mois est environ 14 par tranche de 10,000 habitants pour l'ensemble de l'État. Les taux correspondants pour les populations simulées varient entre 13 et 15.) En ce qui concerne les régions, le biais équivalait aussi à peu près au centième de l'estimation régionale même si cette valeur repose sur un échantillon beaucoup plus petit. Ces résultats donnent à penser que le biais contenu dans nos estimations d'enquête devrait être relativement faible.

Comme le laissait prévoir le faible nombre de comtés dans l'échantillon, la variance de la distribution d'échantillonnage de l'estimateur est assez élevée. L'écart type des estimations obtenues par l'étude de simulation est environ 10 fois plus élevé que le biais. (Pour chacune des cinq simulations, les écarts types du million d'estimations sont du même ordre de grandeur que l'erreur type de l'estimation indiquée dans le tableau 1.) Ces observations permettent de croire que le biais des estimations d'enquête sera plutôt négligeable par rapport à l'erreur type de ces estimations. Finalement, nous avons évalué la forme de la distribution d'échantillonnage de notre estimateur en examinant des histogrammes basés sur le million d'estimations calculées dans chacune des cinq simulations. Les histogrammes paraissaient symétriques, en forme de cloche, et ressemblaient étrangement à des histogrammes de données normales, ce qui nous permet de croire que les intervalles de confiance basés sur l'approximation normale seront assez précis.

4. ESTIMATEUR PAR RÉGRESSION

Des organismes comme le Bureau of the Census mettent à notre disposition de nombreuses données sur la conjoncture économique des comtés. Nous espérons pouvoir utiliser une partie de cette information dans un estimateur par régression afin d'améliorer nos estimations du taux de sans-abri. Nous avons donc cherché à définir un modèle de régression qui mettrait en relation le nombre de sans-abri dans un comté, ou le taux de sans-abri, et divers prédicteurs qui seraient susceptibles, selon nous, d'expliquer la clocharaise. Ces prédicteurs étaient les suivants: population du comté, variation relative de la population entre 1980 et 1990, taux de chômage, proportion des personnes âgées dans la population, dépenses au titre de l'aide sociale, gains hebdomadaires moyens, proportion des maisons de rapport dans le parc immobilier, loyer médian, taux de pauvreté, pourcentage des ménages qui ont à leur tête une femme, pourcentage des terres réservées à l'agriculture, valeur moyenne des exploitations agricoles, revenu moyen par exploitation, rapport des emplois dans le secteur manufacturier aux emplois dans le secteur agricole, indicateur des codes de Beale système de classification ayant pour objet le degré de ruralité (voir Thomas 1977), et indicateurs régionaux.

Une fois les cinq populations obtenues, nous nous sommes servis de chacune d'elles pour évaluer la taille du biais contenu dans les estimations du taux de sans-abri en biais régionaux. Comme nous avions créé une "population" complète, nous pouvions calculer dans les cinq cas le "vrai" taux de sans-abri pour tout l'État et pour les cinq régions géographiques.

Dans la simulation, des échantillons formés de 21 comtés ruraux ont été prélevés selon le plan d'échantillonnage stratifié utilisé dans l'enquête de 1990, c'est-à-dire tirage aléatoire sans remise de quatre comtés dans chacune des régions suivantes: Nord-Est, Nord-Ouest, Centre et Sud-Ouest, et tirage de cinq comtés dans la région Sud-Est. Ensuite, les taux estimés de sans-abri ont été calculés pour les cinq régions et l'État en général au moyen des formules définies dans la section 3.1. Enfin, les estimations ainsi obtenues ont été comparées aux taux de sans-abri pour la population simulée afin d'évaluer le biais contenu dans les estimations. Cette opération en trois étapes (échantillonnage, calcul d'estimations et évaluation du biais) a été répétée un million de fois avec remise pour chaque population simulée. (Le nombre d'échantillons possibles dépasse $7,15 \times 10^{15}$.) La même série de nombres aléatoires a servi au tirage des échantillons pour les cinq populations. Les résultats de la simulation figurent dans le tableau 2.

Tableau 2

Biais contenu dans le taux estimé de sans-abri (nombre de sans-abri par tranche de 10,000 habitants) pour cinq populations simulées (données basées sur 1,000,000 d'échantillons simulés)

Population				
1				
2				
3				
4				
5				

ÉTAT		RÉGION		N-E		N-O		C		S-E		S-O	
0.0406	(2.056)	-0.0406	(3.333)	-0.0578	(3.591)	-0.2442	(3.122)	-0.1034	(6.512)	-0.1034	(4.215)	0.6184	(4.990)
0.1308	(1.759)	-0.0379	(2.923)	-0.2948	(3.194)	0.2700	(3.762)	-0.0279	(4.298)	-0.1132	(4.610)	0.9196	(5.173)
0.2618	(2.144)	0.1538	(3.748)	0.0529	(3.474)	0.3974	(3.426)	-0.1798	(6.600)	-0.1798	(4.274)	1.277	(5.173)
0.2433	(1.807)	0.0317	(4.034)	0.0254	(3.460)	0.1362	(2.260)	0.0427	(3.973)	0.0427	(4.274)	0.6716	(4.274)

Nota: L'écart type de la distribution d'échantillonnage simulée de l'estimateur figure entre parenthèses au-dessous de chaque valeur.

Comme l'estimateur ci-dessus est le rapport de deux variables aléatoires, soit le nombre de sans-abri et l'effectif de grappes échantillonnées, il est entaché d'un biais (voir, par exemple, Cochran (1977)). Le biais diminue à mesure que la taille de l'échantillon (en l'occurrence, le nombre de comités échantillonnés) augmente. Etant donné la faible taille de notre échantillon, nous reconnaissons que nos estimations peuvent être biaisées. Par ailleurs, la fraction de sondage que nous avons utilisée est grande parce que le nombre de comités ruraux est peu élevé. Nous allons donc chercher à évaluer la taille probable du biais contenu dans nos estimations. (Notons que les erreurs types qui figurent dans le tableau I pourraient aussi être inexactes à cause de la faible taille de l'échantillon.)

3.2 Étude de simulation

Nous avons eu recours à une étude de simulation afin d'évaluer la taille probable du biais contenu dans nos estimations. Tout d'abord, nous avons généré cinq "populations" avec, dans chaque cas, le nombre de sans-abri pour chacun des 75 comités non urbains de l'Ohio. Pour les 21 comités qui faisaient partie de l'échantillon, le nombre observé de sans-abri tenait lieu de chiffre officiel; pour les 54 autres comités, les chiffres ont été produits aléatoirement de la manière décrite ci-dessous. Notons que les chiffres simulés sont des chiffres pour la période de six mois.

Nous avons créé la première population en générant le logarithme naturel du taux de sans-abri à partir d'une distribution normale unique. Nous avons choisi le logarithme du taux parce que l'histogramme des taux observés pour les 21 comités de l'échantillon est très asymétrique alors que l'histogramme des logarithmes des taux observés est 2.465, avec un écart type de 0.7154. Nous avons donc généré aléatoirement (à l'aide du logiciel S) les logarithmes des taux de sans-abri à partir d'une distribution normale ayant cette moyenne et cet écart type. Après que les logarithmes ont été générés pour les 54 comités qui ne faisaient pas partie de l'échantillon, ils ont servi, avec les chiffres de population du recensement de 1990 pour chaque comité, à établir le nombre simulé des sans-abri pour les 54 comités.

La deuxième population a été créée de la même manière que la première, sauf qu'il y avait une distribution normale pour chacune des cinq régions géographiques de l'Ohio. Les paramètres de cette distribution, pour chaque région, étaient la moyenne et l'écart type des logarithmes des taux de sans-abri pour les comités de l'échantillon. Là encore, les logarithmes simulés ont servi à établir le nombre de sans-abri pour les 54 comités ruraux non échantillonnés.

La troisième population a été générée au moyen d'une régression du taux de sans-abri (par tranche de 10,000 habitants) par rapport au pourcentage de personnes âgées pour chaque comité de l'échantillon. (Le choix de cette variable explicative est déterminé par l'usage d'un estimateur par régression, comme on l'explique dans la section 4.)

Le modèle de régression ajusté est

$$\widehat{\text{taux}} = -9.02 + 2.32 [\% \text{ de personnes âgées}],$$

avec $R^2 = 0.197$, $\sqrt{\text{EQM}} = 9.03$, et valeur $p = 0.044$ pour le test F global appliqué à la droite de régression. On a créé la population en estimant, pour chaque comité non échantillonné, le taux de sans-abri à partir du pourcentage de personnes âgées, puis en ajoutant un terme d'erreur normale aléatoire. Comme le graphique des résidus de la droite de régression donnait à penser que la variance résiduelle est plus élevée pour les comités où la proportion des personnes âgées est plus forte, on a généré les termes d'erreur aléatoire à partir de deux distributions normales différentes, selon que la proportion des personnes âgées dans un comité était inférieure ou supérieure à 10%. Dans l'un et l'autre cas, c'est l'écart type des résidus qui servait de paramètre pour la distribution normale.

La quatrième population a été générée au moyen d'une régression du taux de sans-abri (par tranche de 10,000 habitants) par rapport au pourcentage de personnes âgées pour chaque comité de l'échantillon et aux indicateurs des régions de l'Etat dans lesquelles se trouvent les comités. Si nous désignons par I_{NE} , I_{NW} , I_C , et I_{SE} les variables indicatrices pour les régions Nord-Est, Nord-Ouest, Centre et Sud-Est respectivement, le modèle de régression ajusté est

$$\widehat{\text{taux}} = -10.40 + 3.23 [\% \text{ de personnes âgées}]$$

$$-6.47I_{NE} - 8.55I_{NW} - 8.64I_C - 14.25I_{SE},$$

avec $R^2 = .407$ (R^2 corrigé = .210), $\sqrt{\text{EQM}} = 8.73$, et valeur $p = 0.127$ pour le test F global appliqué à la droite de régression. On a créé la population en estimant, pour chaque comité non échantillonné, le taux de sans-abri au moyen de l'équation de régression, puis en ajoutant un terme d'erreur normale aléatoire. Là encore, le graphique des résidus de la droite de régression donnait à penser que la variance résiduelle est plus élevée pour les comités où la proportion des personnes âgées est plus forte. On a donc généré les termes d'erreur aléatoire à partir de deux distributions normales différentes, selon que la proportion des personnes âgées dans un comité était inférieure ou supérieure à 10%. Dans l'un et l'autre cas, c'est l'écart type des résidus qui, encore une fois, servait de paramètre pour la distribution normale.

La cinquième population simulée diffère quelque peu de quatre autres. Elle a été générée à l'aide d'un modèle de régression qui permet de prévoir directement le nombre de sans-abri à partir de la taille de la population du comité. Le modèle de régression ajusté est:

$$\widehat{\text{nombre de sans-abri}} = 13.23 + 0.001154 [\text{population}],$$

avec $R^2 = 0.386$, $\sqrt{\text{EQM}} = 54.29$, et valeur $p = 0.003$ pour le test F global appliqué à la droite de régression. On a créé la population en estimant, pour chaque comité non échantillonné, le nombre de sans-abri au moyen de l'équation de régression ajustée, puis en ajoutant un terme d'erreur normale aléatoire. Comme le graphique des résidus donnait à penser que la variance résiduelle est plus

Pour faciliter la comparaison avec les enquêtes qui durent une journée ou une semaine, on a demandé aux sans-abri qui ont participé à notre enquête depuis combien de temps ils étaient dans cette condition. Grâce à cette information, nous avons pu déterminer, pour chaque comité de l'échantillon, le nombre de personnes qui étaient des sans-abri dans la première semaine de l'enquête, soit la première semaine complète de février 1990.

Dans la section 3, nous présentons des estimations du

la première semaine complète de février 1990.

période de six mois. Le taux pour la période de six mois comprend quiconque répondait à la définition du sans-abri à un moment quelconque dans la période d'interview. Le taux pour la période d'une semaine comprend les personnes interviewées au cours des six mois qui disaient avoir été des sans-abri durant la première semaine complète de février.

de février.
Pour éviter de compter un répondant deux fois durant

comme des sans-abri dans les 21 comtés de l'échantillon.

Si les réponses données aux questions filtre par une

de l'information dans le but de connaître les caractéristiques

un questionnaire comportant dix échelles (par exemple, dépression-anxiété, désorientation-déficience de la mémoire, arriération-carence affective) tirées du tableau des états psychiatriques de Spitzler et coll. (1970) (Psychiatric Status Schedule). La troisième section servait à faire le bilan de

Taux estimés de sans-abri dans les régions rurales de l'Ohio
(par tranche de 10,000 habitants)

Nord-Est	3.44 (0.79)	12.00 (2.19)
Nord-Ouest	5.21 (3.51)	12.77 (5.18)
Centre	5.85 (1.86)	12.11 (3.05)
Sud-Est	6.89 (1.93)	15.90 (5.91)
Sud-Ouest	7.25 (2.44)	16.78 (5.32)

où la sommation s'étend aux cinq régions géographiques illustrées dans la figure 1. Les chiffres de population pour les 75 comtés non urbains sont tirés du recensement

$$P_{\text{Etat}} = \frac{\text{population totale des comtés ruraux de l'Ohio}}{\text{population totale des comtés ruraux de l'Ohio}}$$

Le taux estimé de sans-abri pour l'Etat est donc:

a débuté avec la première semaine complète de février 1990. Dans chaque comité de l'échantillon, un réseau d'ajustage s'occupait d'identifier et de trouver les sans-abri. Chaque réseau était sous la surveillance d'un coordonnateur local. Des enquêteurs principaux surveillaient le travail du coordonnateur de comité et des employés du bureau central. Par des appels téléphoniques aux deux semaines et des visites sur le terrain, les enquêteurs principaux contrôlaient la collecte des données afin d'assurer l'uniformité des procédures et la qualité des opérations.

Les conseillers et les intervieweurs, qui avaient été choisis à cause de leur bonne connaissance du comité dans lequel ils travaillaient, s'occupaient d'identifier les gens qui répondaient à la définition d'un sans-abri. Parmi les conseillers, on retrouvait des pasteurs, des employés d'hôpital, des dirigeants d'organismes de bienfaisance, des représentants élus, des animateurs de la communauté, des barmans, des commis d'hôtel, des employés de laverie automatique, de même que des fournisseurs de services professionnels tels que les membres des départements de santé communautaire, les bibliothécaires, les conseillers agricoles, les employés des postes, les ministres du culte, les rangers, les groupes d'action communautaire, les travailleurs sociaux de cas, les travailleurs de santé mentale et les agents de la paix. Quant aux intervieweurs, ils étaient une centaine à réaliser les interviews auprès des sans-abri, après avoir assisté à une séance de formation de quatre heures et s'être vu remettre un guide des opérations sur le terrain. Les interviews se sont déroulées dans des bureaux, des petits restaurants, des chambres de motel, des automobiles, des parcs publics, des laveries automobiles, des bars et sous des viaducs de chemin de fer. Les intervieweurs avaient été informés des ressources communautaires existantes et avaient la compétence nécessaires pour orienter éventuellement des répondants vers les services appropriés. De plus, les intervieweurs disposaient d'un certain budget au cas où ils seraient appelés à dépanner un répondant (moins de 600 \$ ont été dépensés de cette manière). Il fallait limiter l'aide offerte par l'intermédiaire des intervieweurs pour ne pas inciter des gens à se faire passer pour des sans-abri.

2.2 Définition de la clochardisie en région rurale

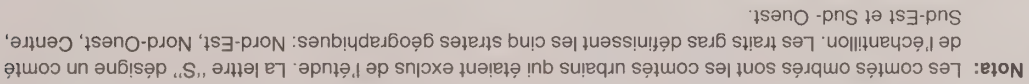
On s'est servi de questions filtrant pour identifier les sans-abri. La définition de la clochardisie utilisée pour cette étude différait nécessairement quelque peu de celle qui avait servi pour les études en région urbaine. En effet, il existe moins de lieux de refuge pour les itinérants dans les régions rurales. Un répondant était reconnu comme un sans-abri s'il prétendait ne pas avoir de domicile fixe et si, la nuit précédant l'interview, il avait dormi soit 1) à la belle étoile, soit 2) dans un refuge pour sans-abri, soit 3) dans un hôtel ou un motel bon marché, pourvu que la durée de séjour réelle ou prévue fût de 45 jours ou moins, ou 4) dans un autre lieu particulier, pourvu que la durée de séjour réelle ou prévue fût de 45 jours ou moins. La quatrième catégorie comprenait les gens qui séjournaient dans des hangars, des granges, de vieux autobus ou de vieilles roulotte sans eau

2.3 Période d'interview

ni électricité, pourvu qu'ils n'en étaient pas les propriétaires et qu'ils ne payaient pas de loyer pour demeurer là. Étaient reconnues aussi comme des sans-abri les personnes qui résidaient temporairement chez des amis ou des parents, pourvu qu'elles étaient arrivées depuis moins de 46 jours, qu'elles ne faisaient pas partie à proprement parler du ménage et qu'elles prévoyaient quitter dans un délai de 45 jours. Les personnes qui séjournaient dans des refuges pour femmes battues, des hôpitaux, des prisons, des camps de travailleurs migrants, etc. n'étaient pas comptées parmi les sans-abri à moins qu'elles n'eussent aucun endroit où aller au moment de quitter l'établissement.

On peut comparer notre définition de la clochardisie avec celle utilisée dans les études sur les sans-abri en région urbaine. C'est dans la Stewart B. McKinney Homeless Assistance Act (1987) que l'on trouve les éléments de la définition de la clochardisie au sens des études sur les sans-abri en région urbaine. Selon cette loi, un ou une sans-abri est "une personne qui n'a pas d'endroit fixe, régulier et convenable où passer la nuit et qui doit compter surtout sur 1) les établissements supervisés à caractère public ou privé qui ont pour mission d'offrir des services d'hébergement temporaire (c.-à-d. établissements de bien-être, habitations collectives et logements de transition pour les personnes souffrant de troubles psychiques); 2) les institutions qui offrent des services d'hébergement temporaire aux personnes qui doivent être placées en établissement; ou 3) des endroits publics ou privés qui ne sont pas considérés normalement comme des lieux d'hébergement pour la nuit" (TRADUCTION). Comme le laisse entendre l'article de Rossi et coll. (1987), de cette définition découle la notion de "sans-abri au sens littéral". La définition ci-dessus ne recouvre pas, par exemple, les sans-abri qui demeurent temporairement chez des amis ou des parents. Or, nous avons choisi de tenir compte de cette catégorie de sans-abri dans notre étude. Selon notre analyse, environ le tiers des personnes qui ont été dénombrées lors de notre enquête la clochardisie pour les régions urbaines. On ne peut dire de combien augmenterait le chiffre estimé des sans-abri dans les régions urbaines si on tenait compte des personnes qui demeurent temporairement chez des amis ou des parents.

La période de six mois utilisée pour dénombrer les sans-abri dans les régions rurales tranche avec la période d'une journée qui est utilisée normalement dans les enquêtes faites en région urbaine. Dans une analyse de sept enquêtes ayant pour objet les sans-abri, Burt et Tauber (1991) remarquent que la période d'interview pouvait correspondre à une seule journée (en soirée notamment) ou pouvait s'étendre sur une ou deux semaines et les interviews avaient lieu à un seul endroit. Dans la plupart de ces enquêtes, on cherchait les sans-abri à interviewer dans les refuges, les soupes populaires, les bâtiments désaffectés ou les autres endroits semblables. Comme, en région rurale, les refuges pour sans-abri et les soupes populaires sont plus rares, il est plus difficile de trouver les sans-abri, ce pourquoi une période d'enquête plus longue est recommandée.



Le plan de sondage qui a servi à l'enquête de 1990 a été choisi expressément pour faciliter la comparaison avec une étude qui a été réalisée en 1984 sur les sans-abri des régions rurales de l'Ohio (Roth et coll. 1985). Dans cette étude, on avait groupé les comtés de l'Ohio en cinq régions (Nord-Est, Nord-Ouest, Centre, Sud-Est et Sud-Ouest) et on avait prélevé 16 comtés ruraux au moyen d'un plan d'échantillonnage aléatoire stratifié. Les 21 comtés qui ont été échantillonnés pour l'enquête de 1990 comprenaient les 16 comtés de l'étude de 1984 plus un comté tiré au hasard dans chacune des cinq régions. (Notons que l'analyse des données de l'enquête de 1990 nous permet de croire que la stratification du territoire de l'Ohio ne contribue pas à améliorer l'estimation du taux de sans-abri en région rurale.) Dans la figure 1, on peut voir une carte

Nous avons tenté de recenser tous les sans-abri des 21 comités de l'échantillon. Comme il n'existe pas nécessairement de lieux de refuge ou de rencontre pour les sans-abri dans les régions non urbaines, l'enquête s'est étalée sur une période de six mois et a fait appel à un réseau de conseillers pour trouver les sans-abri. La période d'enquête

de l'Ohio où sont indiqués les cinq régions, les comtés urbains et les comtés qui font partie de l'échantillon. Dans les sous-sections suivantes, nous décrivons brièvement la méthodologie de l'enquête de 1990. Pour une description plus complète, prière de consulter First et coll. (1993) ainsi que Toomey et coll. (1993).

Estimation du taux de sans-abri dans les régions rurales: une étude des régions non urbaines de l'Ohio

ELIZABETH A. STASNY, BEVERLY G. TOOMEY et RICHARD J. FIRST¹

RÉSUMÉ

Depuis quelques temps, on consacre beaucoup d'effort pour dénombrer les sans-abri et en définir les caractéristiques. Cependant, on s'est concentré jusqu'à maintenant sur les sans-abri des régions urbaines. Dans cet article, nous décrivons les efforts qui ont été faits en vue d'estimer le taux de sans-abri dans les comtés non urbains de l'Ohio. Les méthodes qui servent à repérer les sans-abri et même la définition de la clochardisie sont différentes dans les régions rurales, où les institutions qui accueillent les sans-abri sont moins nombreuses. De plus, l'emploi d'estimateurs ordinaires basés sur des échantillons d'enquête peut poser des difficultés dans une analyse axée sur les régions rurales, étant donné que ces estimateurs exigent généralement de grandes populations, de grands échantillons et de faibles fractions de sondage. Nous décrivons l'enquête qui a été faite dans les régions non urbaines de l'Ohio pour dénombrer les sans-abri et nous présentons l'étude de simulation qui a été effectuée dans le but d'évaluer l'utilité des estimateurs ordinaires de la proportion d'une population basés sur un échantillon en grappes stratifié.

MOTS CLÉS: Estimateur biaisé; estimateur par régression; petit échantillon; échantillon en grappes stratifié; simulation.

1. INTRODUCTION

La notion de sans-abri évoque souvent dans notre esprit l'idée des "marginaux de rues" et des "clochardes". Nous nous représentons des personnes qui dorment sur des bancs de parc ou des grilles à air ou dans des lieux de refuge pour sans-abri. Cependant, ces images stéréotypées viennent des grandes villes et ne décrivent pas nécessairement avec exactitude les conditions de vie des sans-abri en région rurale. Bon nombre des études portant sur les sans-abri ont été réalisées dans les grands centres urbains. Par exemple, l'étude de 1987 du Urban Institute visait à dénombrer les sans-abri de 20 grandes villes américaines. Rossi a effectué une autre grande étude du même genre à Chicago. (Pour avoir un aperçu des méthodes d'enquête utilisées dans ces études et d'autres études qui ont servi à dénombrer des populations de sans-abri, prière de se référer à Burt et Tauber (1991).)

L'occasion du recensement de la population de 1990 aux États-Unis, on a tenté de dénombrer les sans-abri par l'opération S-Night (Shelter and Street Night). On avait dressé spécialement à cette fin une liste des lieux de refuge des sans-abri à la grande échelle du pays. Le plus haut dirigeant élu de plus de 39,000 municipalités rurales et urbaines devait fournir une liste des endroits (refuges, lieux publics, etc.) où les sans-abri passent la nuit. Les recenseurs ont procédé au dénombrement des sans-abri dans une seule nuit, celle du 20 mars. Il convient de souligner que l'opération S-Night avait pour but essentiellement de dénombrer les sans-abri; les données du recensement contiennent relativement peu d'information sur les caractéristiques des sans-abri. Pour plus de détails sur l'opération S-Night, prière de se référer à Tauber et Siegel (1990).

Le territoire de l'Ohio est divisé en 88 comtés, dont 13 sont des comtés urbains qui comprennent de grandes villes et 75 sont définis comme des comtés ruraux ou non urbains. Parmi ces 75 comtés, qui font l'objet de notre étude, il y a des comtés entièrement ruraux, des comtés qui ne sont pas voisins de comtés urbains et qui ont un chef-lieu moyennement peuplé, ainsi que des comtés suburbains (c.-à-d. voisins de comtés dans lesquels se trouvent de grandes régions métropolitaines).

2. DESCRIPTION DE L'ENQUÊTE

Contrairement aux enquêtes faites dans les régions urbaines et à l'opération S-Night, l'enquête que nous décrivons ici avait pour but de repérer et de dénombrer les sans-abri des régions non urbaines, où qu'ils soient, et de recueillir de l'information qui permettrait de définir les caractéristiques de ces personnes. Dans la section qui suit, nous décrivons le plan de sondage de l'enquête de 1990 sur les sans-abri des régions rurales de l'Ohio. Nous donnons notre définition de la clochardisie en région rurale et nous décrivons les méthodes utilisées pour repérer et interviewer les sans-abri. Dans la section 3, nous présentons les estimations du taux de sans-abri en région rurale que nous avons calculées à l'aide de l'estimateur ordinaire de la proportion basé sur un échantillon en grappes stratifié. Comme ces estimations sont susceptibles d'être biaisées, nous présentons aussi les résultats d'une étude de simulation qui avait pour but d'évaluer la taille probable du biais. Dans la section 4, nous considérons un estimateur par régression pour calculer le taux de sans-abri et nous comparons cet estimateur à celui utilisé dans la section 3. Enfin, la section 5 renferme les conclusions de notre étude.

¹ Elizabeth A. Stasny, Department of Statistics; Beverly G. Toomey et Richard J. First, College of Social Work, The Ohio State University, Columbus, Ohio 43210.

- COCHRAN, W.G. (1942). Sampling theory when the units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3ième Ed. New York: John Wiley.
- COPELAND, K.R., PEITZMEIER, F.K., et HOY, C.E. (1987). Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population. *Techniques d'enquête*, 13, 183-191.
- COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- COX, L.H., et ERNST, L.R. (1982). Controlled rounding. *INFOR*, 20, 423-432.
- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEMING, W.E., et STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- DARROCH, J.N., et RATCLIFF, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43, 1470-1480.
- EL-BADRY, M.A., et STEPHAN, F.F. (1985). On adjusting sample tabulations to census counts. *Journal of the American Statistical Association*, 50, 738-762.
- FAGAN, J.T., GREENBERG, B.V., et HEMMIG, B. (1988). Controlled rounding of three dimensional tables. *Statistical Research Division Report Census/SRD/RR-88/02*. U.S. Bureau of the Census, Washington, D.C.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C*, 37, 117-132.
- FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., et PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames Iowa.
- GOEBEL, J.J. (1976). Application of an iterative regression technique to a national potential cropland survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 350-353.
- HAMPPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the Statistical Computing Section, American Statistical Association*, 59-64.
- HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat non-publiée, Iowa State University, Ames, Iowa.
- HIDIROGLOU, M.A., FULLER, W.A., et HICKMAN, R.D. (1976). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.
- HUANG, E.T. (1978). Nonnegative regression estimation for sample survey data. Thèse de doctorat non-publiée, Iowa State University, Ames, Iowa.
- HUANG, E.T. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section, American Statistical Association* 1978. 300-303.
- HULLIGER, B. (1993). Robustification of the Horvitz-Thompson estimator. Document présenté à la 49ième session de l'Institut International de Statistique. Livre 1, 583-584.
- HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Thèse de maîtrise non-publiée, Iowa State University, Ames, Iowa.
- IRELAND, C.T., et KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 169-188.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Experiment Station Research Bulletin, 304.
- KRASAKER, W.A. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, 48, 1333-1346.
- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.
- LUERY, D. (1986). Weighing survey data under linear constraints on the weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 325-330.
- MALLOWS, C.L. (1983). Discussion of Huber: Minimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78, 77.
- MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.
- OH, H.L., et SCHUBERT, F. (1987). Variante de la méthode itérative du quotient. *Techniques d'enquête*, 13, 221-232.
- RAO, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. Document présenté au Workshop on Users of Auxiliary Information in Surveys, Örebro, Sweden, Octobre, 1992.
- ROYAL, R.M., et CUMBERLAND, W.G. (1981). The finite-population linear regression estimator and estimators of its variance – an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SAS INSTITUTE INC. (1989). SAS/STAT User's Guide, Version 6 Fourth Edition, Volume 1. Cary, NC: SAS Institute Inc.
- SINGH, A.C. (1993). On weight adjustment in survey sampling. Document non publié. Statistique Canada, Ottawa, Canada.
- STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.
- WATSON, D.J. (1937). The estimation of leaf areas. *Journal of Agricultural Science*, 27, 474-483.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.
- ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

$$J_n = (1, 1, \dots, 1)',$$

$$A^{(0)} = Z'H^{(0)}Z,$$

$$G^{(0)} = \text{diag}(g_1^{(0)}, \dots, g_n^{(0)})$$

et

$$H^{(0)} = VG^{(0)}.$$

L'algorithme consiste dans l'itération de trois opérations.

1. Le calcul initial est pour $\alpha = 0$. À l'itération α , le

vecteur des poids de régression, désigné par $w^{(\alpha)}$, est

$$w^{(\alpha)} = [1 + nu^{(\alpha)}]^{-1} V(n^{-1}J_n + n^{(\alpha)})$$

$$= (w_1^{(\alpha)}, \dots, w_n^{(\alpha)})', \quad (\text{A.5})$$

où

$$n^{(\alpha)} = G^{(\alpha)}Z(A^{(\alpha)})^+(X - X^v) = (n_1^{(\alpha)}, \dots, n_n^{(\alpha)}),$$

$$X^v = \left(\sum_{i=1}^n v_i \right) \left(\sum_{i=1}^n v_i x_{iv} \right)$$

($A^{(\alpha)}$)[†] est un inverse généralisé symétrique de $A^{(0)}$,

$$nu^{(\alpha)} = \max\{J_n' V n^{(\alpha)}, n^{-1} - 1\}, \quad (\text{A.6})$$

et

$$A^{(\alpha)} = Z'H^{(\alpha)}Z.$$

2. Le programme vérifie si les poids obtenus à l'étape 1 satisfont les critères.

(a) Est-ce que $|nu^{(\alpha)}| \leq M$ pour tous i ?

(b) Est-ce que

$$L_B \leq nw_i^{(\alpha)} \leq U_B$$

pour tous i ?

Si le critère (a) ou (b) n'est pas respecté pour i importe quel i et que le programme n'a pas complété $L1$ itérations, on passe à l'étape 3. Si les critères (a) et (b) sont satisfaits ou si le programme a complété $L1$ itérations, les poids obtenus à cette étape sont produits pour analyse.

3. Les facteurs de contrôle $h_i^{(\alpha)}$, $i = 1, 2, \dots, n$, sont modifiés. Posons

$$H^{(\alpha)} = H^{(\alpha-1)}G^{(\alpha)},$$

On passe à l'étape 1 pour calculer de nouveaux poids de régression.

La constante 1.33 dans la définition de $d_i^{(\alpha)}$ et la constante 0.8 dans la définition de $g_i^{(\alpha)}$ ont été choisies pour accélérer la convergence. Les facteurs de contrôle $g_i^{(\alpha)}$ servent à sous-pondérer les observations suivant une distance par rapport à la moyenne de population.

La définition de $w^{(\alpha)}$ en (A.5) représente une autre manière d'écrire le vecteur des poids par les moindres carrés généralisés définis en (2.4) lorsque $\pi_{i-1}^* = h_i^{(\alpha)}$.

BIBLIOGRAPHIE

- AKKERBOOM, J.C., SIKKEL, D., et van HERK, H. (1991). Robust weighting of financial survey data. Article présenté à la réunion de l'Institut International de Statistique, Cairo, Egypt.
- ALEXANDER, C.H. (1987). A model based justification for survey weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.
- BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- BETHLEHEM, J.G., et KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- BRACKSTONE, G.J., et RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā, Series C*, 97-114.
- BUREAU OF THE CENSUS (1987). Current Population Survey, March 1987; Technical Documentation. Washington, D.C.

régression. Or, cette dernière ne représente qu'environ 6% de la population et on observe peu de corrélation entre les écarts par rapport à la moyenne de sous-population et les variables de contrôle. Par ailleurs, on considère que les estimations par régression sont beaucoup plus efficaces que les estimations non pondérées pour la sous-population des ménages qui tiennent la maison. L'efficacité relative des estimations par régression pour cette sous-population se rapproche de celle des estimations pour l'ensemble de la population.

Même si l'on tient compte du fait que les chiffres de population tirés de la Current Population Survey ne sont pas des chiffres officiels, il est clair que l'estimation par régression de moyennes de population engendre des gains d'efficacité appréciables. Et, bien que l'on considère qu'un estimateur par régression de la moyenne d'une petite sous-population est moins efficace qu'un estimateur non pondéré du même paramètre, la perte d'efficacité est faible par rapport aux gains d'efficacité notables estimés pour les autres variables.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à un contrat de soutien à la recherche (n° 58-3198-9-032) passé avec le Human Nutrition Information Service du Département de l'Agriculture des E.-U. Les auteurs remercient Phil Kott, Patricia Guenther et les arbitres pour leurs commentaires utiles.

ANNEXE

PROGRAMME DE PRODUCTION DE POIDS

Dans cette annexe, nous décrivons le programme de production de poids de régression de Huang et Fuller (1978). Ce programme offre la possibilité de définir un poids minimum et un poids maximum. La version originale du programme n'offrait pas cette possibilité. Pour une analyse de méthodes de production de poids connexes, voir Singh (1993).

Supposons connues les moyennes de population (X_1, X_2, \dots, X_k) des k variables auxiliaires (X_1, X_2, \dots, X_k). Supposons aussi qu'il existe un échantillon de n observations et posons

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}, \quad (A.1)$$

où X_{ij} est la valeur observée de la variable j pour l'individu i .

Outre le tableau d'observations d'échantillon et les moyennes de population, deux facteurs initiaux, v_i et $g_i^{(0)}$, $i = 1, 2, \dots, n$, sont nécessaires au calcul. Le

facteur v_i est en général inversement proportionnel à la probabilité de sélection. La valeur implicite de $g_i^{(0)}$ est l'unité (c.-à-d. $g_i^{(0)} = 1$). Dans le cas d'échantillons stratifiés ou de données avec variances inégales, l'utilisateur peut choisir d'autres valeurs pour $g_i^{(0)}$. (Voir Huang 1978 ou Goebel 1976.) Les paramètres du programme sont la taille de l'échantillon (n), la taille de la population (N), le paramètre M , le nombre maximum d'itérations (L_1), la limite supérieure du rapport du poids au poids moyen (L_B). Il faut que $0 \leq L_B < 1 < U_B$. Dans notre description, nous supposons que $\sum_{i=1}^n v_i = n$. Le programme normalise les v_i de sorte que leur somme soit égale à n .

Le programme peut servir à créer des poids en vue de l'estimation de moyennes ou de totaux. Les poids pour les totaux équivalent aux poids pour les moyennes multipliés par N . En ce qui concerne les moyennes, le programme vise à créer des poids w_i qui soient tels que

$$\sum_{i=1}^n w_i(1, X_i) = (1, X), \quad (A.2)$$

$$L_B < n w_i < U_B, \quad (A.3)$$

$$(1 - M) \max_{1 \leq i \leq n} w_i v_i \leq (1 + M) \min_{1 \leq i \leq n} w_i v_i, \quad (A.4)$$

pour $i = 1, 2, \dots, n$.

Le programme est itératif. En l'occurrence, une itération consiste à calculer les poids par les moindres carrés généralisés, un facteur de contrôle h_i étant appliqué à chaque observation. Le facteur h_i est le produit de v_i par g_i , g_i étant, après la première itération, une fonction – décrite par une courbe en cloche – de la distance (exprimée en une mesure appropriée) entre l'observation et la moyenne de population. À chaque itération, les poids satisfont l'équation (A.2) mais peuvent ne pas satisfaire (A.3) ou (A.4). Il ne sera pas toujours possible de créer des poids qui satisfont les conditions spécifiées dans le nombre d'itérations voulu. Si l'échantillon est tel qu'il n'est pas possible de respecter la contrainte, le programme produit les poids de la dernière itération. Dans le cas d'une variable x unique, si le critère ne peut être respecté, il y aura deux poids: un pour les valeurs supérieures à la moyenne de population et un autre pour les valeurs qui lui sont inférieures.

Pour décrire l'algorithme, posons

$$Z_{ij} = X_{ij} - \bar{X}_j,$$

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \dots & Z_{np} \end{pmatrix},$$

$$V = \text{diag}(v_1, v_2, \dots, v_n),$$

Les valeurs estimées des coefficients de corrélation multiple cartée entre les variables du tableau 2 et les 27 variables de contrôle sont de 0.29 pour les repas à l'extérieur, de 0.44 pour les aliments à la maison, de 0.82 pour les repas-personnes et de 0.12 pour la tenue de la maison. Si les moyennes d'échantillon des variables de contrôle étaient presque égales aux moyennes de population, l'erreur type de l'estimation par régression de la moyenne pour les repas à l'extérieur serait d'environ 84 centièmes $((1 - 0.29)^{1/2})$ de l'erreur type de l'estimation non pondérée. En fait, l'erreur type estimée de l'estimation par régression équivalait à peu près à 97 centièmes de l'erreur type de l'estimation non pondérée. L'écart est dû au fait que $\sum_{i=1}^n w_i^2$ est beaucoup plus grande que n^{-1} parce que l'échantillon accuse un déséquilibre pour un certain nombre de caractéristiques. Notons que

$$0.97 \approx [(0.71)(1.32)]^{1/2},$$

où $0.71 = (1 - 0.29)$ est un moins le carré du coefficient de corrélation et $1.32 = n \sum_{i=1}^n w_i^2$. En ce qui concerne la variable "tenue de la maison", la situation est plus particulière. La corrélation n'est pas forte et, semble-t-il, les écarts par rapport à la droite de régression les plus prononcés sont associés à des poids élevés. La variance estimée de l'estimateur par régression est d'environ le double de celle de l'estimateur non pondéré.

Le tableau 2 contient aussi la valeur estimée de l'écart entre l'estimation non pondérée et l'estimation pondérée. L'écart entre le total estimé non pondéré et le total estimé pondéré est

$$\sum_{i=1}^n N n_{i-1} y_i - \sum_{i=1}^n w_i y_i = \sum_{i=1}^n (n_{i-1} N - w_i) y_i.$$

L'écart entre les moyennes estimées est égal au quotient de la différence de totaux par la taille de la population. Pour calculer la variance de l'écart entre les moyennes, notons que l'hypothèse d'un écart nul équivaut à l'hypothèse d'une corrélation nulle entre w_i et y_i . Nous avons donc effectué la régression non pondérée de y_i par rapport à w_i et nous avons calculé la variance du coefficient de régression selon le plan au moyen de PC CARP. Les erreurs types des écarts estimés du tableau 2 sont telles que la "statistique t " pour l'hypothèse de l'écart nul est égale à la "statistique t " pour le coefficient de w_i dans la régression de y_i par rapport à w_i .

Pour les quatre caractéristiques, l'écart entre l'estimation pondérée et l'estimation non pondérée de la moyenne de population est significatif aux niveaux habituels. Par conséquent, si on suppose que les estimateurs par régression sont non biaisés, les estimateurs non pondérés, eux, comportent un biais appréciable. En ce qui regarde les estimations des moyennes de sous-population, les résultats sont partagés. L'écart entre l'estimation pondérée et l'estimation non pondérée est significatif pour les trois moyennes relatives à la sous-population des ménages tenant la maison – qui est en fait la population étudiée – alors qu'il ne l'est pas pour les trois

moyennes relatives à la sous-population des ménages qui ne tiennent pas la maison. L'échantillon ne compte que 222 ménages de ce genre. Par conséquent, la variance de l'écart entre l'estimation pondérée et l'estimation non pondérée est beaucoup plus grande pour les ménages qui ne tiennent pas la maison que pour les autres. L'écart entre les deux estimations de la moyenne de population est fonction de l'écart entre les deux estimations de la moyenne de sous-population et de l'écart entre les deux estimations de la proportion de ménages de l'une et l'autre catégories. Voilà pourquoi l'écart pour l'"ensemble de la population" peut être plus grand que l'écart pour l'une et l'autre sous-populations. La dernière colonne du tableau 2 contient le rapport de l'erreur quadratique moyenne estimée de l'estimateur non pondéré à la variance de l'estimateur par régression. L'erreur quadratique moyenne estimée de l'estimateur non pondéré a été calculée par la formule

$$\widehat{\text{EQM}}_u = V + \max\{0, (\text{Diff})^2 - (e.l. \text{diff})^2\},$$

où V est la variance estimée de l'estimation non pondérée, "Diff", l'écart entre les deux estimations dans le tableau 2, et "e.l. diff", l'erreur type de l'écart dans le tableau 2. La variance estimée V de l'estimateur non pondéré est définie par la formule (2.8), où w_{qit} est constant et où $x_{qit} \beta$ est remplacé par y_{it} . Le second terme de l'erreur quadratique moyenne estimée est le carré du biais estimé. Suivant l'hypothèse que l'estimateur par régression est non biaisé, l'espérance de $(\text{Diff})^2$ est égale à la somme du carré du biais et de la variance de l'écart. Donc, suivant l'hypothèse que l'estimateur par régression est non biaisé, l'espérance mathématique de $(\text{Diff})^2$ est le carré du biais plus la variance de la différence. L'erreur quadratique moyenne estimée de l'estimateur pondéré correspond à la variance de l'estimateur pondéré calculée comme le carré de l'erreur type dans le tableau 2.

Si l'on considère les quatre caractéristiques pour lesquelles des moyennes de population ont été estimées, l'efficacité relative estimée de l'estimateur par régression (par rapport à un estimateur simple) varie de 2.5 à 129. La valeur la moins élevée se rapporte à la variable "repas à l'extérieur". Les variances des deux estimateurs sont semblables mais, à cause du biais estimé, l'erreur quadratique moyenne (EQM) de l'estimation par régression pour cette variable équivaut à environ 40% de l'EQM de l'estimation non pondérée. Pour la variable "aliments à la maison", l'EQM de l'estimation par régression équivaut à moins de 20% de celle de l'estimation non pondérée; pour les deux autres variables, le rapport correspondant est d'environ 1% (repas-personnes) et 20% (tenue de la maison). Dans tous les cas, le carré du biais est une composante très importante de l'erreur quadratique moyenne estimée. Comme on observe un biais relativement faible dans les estimations non pondérées touchant la sous-population des ménages qui ne tiennent pas la maison, on considère que pour cette sous-population les estimations non pondérées sont un peu plus efficaces que les estimations par

Tableau 2
Propriétés de divers estimateurs

Variable	Moyenne non pondérée	Moyenne pondérée	Écart	Efficacité relative de la régression	
Repas à l'extérieur	Ménages tenant	7.75	7.93	-0.18	2.52
	maison	(0.22)	(0.17)	(0.09)	
	Ménages ne tenant	18.31	18.12	0.19	0.92
	pas maison	(1.14)	(1.19)	(0.68)	
	Ensemble de la population	8.27	8.57	-0.30	2.56
Aliments à la maison	Ménages tenant	61.10	59.56	1.54	3.65
	maison	(1.14)	(0.98)	(0.41)	
	Ménages ne tenant	25.99	26.39	-0.40	0.73
	pas maison	(1.25)	(1.46)	(1.00)	
	Ensemble de la population	59.37	57.49	1.88	5.60
Repas-personnes	Ménages tenant	2.42	2.33	0.09	89.00
	maison	(0.03)	(0.01)	(0.01)	
	Ménages ne tenant	0.51	0.49	0.02	1.00
	pas maison	(0.03)	(0.03)	(0.02)	
	Ensemble de la population	2.33	2.22	0.11	129.00
Tenue de la maison (%)	Ménages tenant	95.06	93.77	1.29	5.30
	maison	(0.40)	(0.58)	(0.10)	
	Ménages ne tenant	0.49	0.47	0.02	
	pas maison	(0.03)	(0.03)	(0.02)	
	Ensemble de la population	2.33	2.22	0.11	129.00

estimation pour un plan donné et la variance de l'estimation pour un échantillon aléatoire simple de même taille est appelé "effet du plan". En ce qui concerne les moyennes estimées à l'aide de données non pondérées pour l'ensemble de la population, l'effet du plan estimé est d'environ 2.5 pour les repas à l'extérieur et les repas-personnes, d'environ 4.1 pour les aliments à la maison et d'environ 1.5 pour la tenue de la maison.

La colonne intitulée "moyenne pondérée" contient les estimations calculées au moyen des poids de régression. Les erreurs types correspondantes ont été calculées sur PC CARP au moyen de la formule (2.8), les π_i^{-1} étant remplacés par les poids de régression. Pour calculer la variance y . Les moyennes estimées pour les sous-populations sont des rapports d'estimateurs par régression. On a calculé la variance de ces moyennes estimées en calculant au moyen de la formule (2.8) la variance des écarts aléatoires de Taylor pour les rapports en question. Les estimations non pondérées et les estimations pondérées ont des erreurs types semblables en ce qui concerne les repas à l'extérieur et les aliments à la maison. Cependant, pour les repas-personnes, l'erreur type de l'estimation par régression de la population est d'environ un tiers de l'erreur type de l'estimation non pondérée. L'erreur type de l'estimateur par régression est moins élevée parce que la variable "repas-personnes" est fortement corrélée avec les variables de taille du ménage utilisées comme variables de contrôle dans la régression.

repas-personnes pour le j -ième ménage

$$= \sum_i (h_{ij} + a_{ij})^{-1} h_{ij} + (21)^{-1} b_j,$$

où h_{ij} = nombre de repas pris à la maison par l'individu i dans le ménage j durant la semaine d'interview, a_{ij} = nombre de repas pris à l'extérieur du foyer par l'individu i dans le ménage j durant la semaine d'interview et b_j = nombre de repas servis à des personnes qui ne sont pas membres du ménage dans le ménage j durant la semaine d'interview.

Le nombre total redressé de repas pris à l'extérieur du foyer est la somme des proportions des repas pris à l'extérieur par chaque membre du ménage durant la semaine d'interview, multipliée par 21. En reprenant la notation utilisée plus haut pour les repas-personnes, nous avons

repas à l'extérieur pour le j -ième ménage

$$= 21 \sum_i (h_{ij} + a_{ij})^{-1} a_{ij}.$$

La valeur totale des aliments consommés à la maison est égale à la somme des dépenses d'alimentation et de la valeur monétaire des aliments produits à la maison et des aliments reçus gratuits qui ont été consommés durant la semaine d'enquête. Les dépenses d'alimentation ont été évaluées au prix que l'on disait avoir payé, abstraction faite du moment de l'achat et la taxe de vente exclue. En ce qui concerne les aliments dont le prix n'était pas déclaré, les aliments produits à la maison et les aliments reçus en cadeau ou en guise de paiement, ils ont été évalués au prix moyen la livre qu'avaient payé les ménages visés par l'enquête pour de la nourriture semblable dans la même région et à la même période de l'année.

Un ménage qui tient la maison est un ménage où au moins une personne a un nombre redressé de dix repas ou plus avec les réserves alimentaires du ménage dans les sept jours précédant l'interview. En règle générale, les études de consommation alimentaire des ménages portent uniquement sur des ménages qui tiennent la maison.

Les moyennes de variables calculées à l'aide de données non pondérées figurent dans la colonne "moyenne non pondérée" du tableau 2. Trois moyennes sont calculées pour chacune des trois variables suivantes: "repas à l'extérieur", "aliments à la maison" et "repas-personnes". Deux moyennes sont calculées pour les deux sous-populations définies par la variable indicatrice "tenue de la maison". La troisième, qui correspond à la rubrique "ensemble de la population", est la moyenne estimée pour la population totale. L'erreur-type de l'estimation figure entre parenthèses sous chaque valeur estimée. Les estimations non pondérées et les erreurs types correspondantes ont été calculées à l'aide de PC CARP; voir Fuller et coll. (1986). Le calcul a tenu compte du fait qu'il s'agissait d'un échantillon en grappes stratifié aréolaire. Comme il s'agit d'un échantillon à deux degrés, les variances estimées sont plus grandes que la variance d'un échantillon aléatoire simple qui contiendrait le même nombre de ménages. Le rapport entre la variance d'une

printemps et environ 16% dans chaque cas au trimestre d'été et au trimestre d'automne. Des interviews ont eu lieu aux trimestres de printemps et d'été de 1987 et 1988. L'échantillon n'était pas non plus équilibré en ce qui a trait à la caractéristique "degré d'urbanisation". Les ménages vivant dans les noyaux urbains étaient sous-représentés dans l'échantillon (24%, contre 31% dans la population), alors que les ménages vivant dans les régions non métropolitaines étaient surreprésentés (29%, contre 23% dans la population). Les ménages à revenu élevé étaient sous-représentés alors que c'était le contraire pour les ménages ayant à leur tête et un homme et une femme (68%, contre 61% dans la population). Les ménages avec enfants étaient, eux, surreprésentés. Pour plusieurs autres caractéristiques socio-démographiques, le déséquilibre était assez peu prononcé.

Le personnel du Human Nutrition Information Service est d'avis que les caractéristiques énumérées dans le tableau 1 ont un rapport avec les habitudes de consommation alimentaire. C'est pourquoi on s'est servi de variables basées sur ces caractéristiques dans la pondération par régression. Pour mettre en marche le programme de production de poids, on a transformé chacune des variables nominales du tableau 1 en un ensemble de variables indicatrices. Par exemple, on a créé trois variables pour la caractéristique "revenu du ménage en pourcentage du seuil de pauvreté". Ces trois variables sont les suivantes:

$$\begin{aligned} Z_{i1} &= 1 \quad \text{si la proportion pour le } i\text{-ième ménage est inférieure à } 131\% \\ &= 0 \quad \text{autrement,} \\ Z_{i2} &= 1 \quad \text{si la proportion pour le } i\text{-ième ménage se situe entre } 131 \text{ et } 300\% \\ &= 0 \quad \text{autrement,} \\ Z_{i3} &= 1 \quad \text{si la proportion pour le } i\text{-ième ménage se situe entre } 301 \text{ et } 500\% \\ &= 0 \quad \text{autrement.} \end{aligned}$$

Vingt-cinq variables indicatrices ont été créées de cette manière. En outre, la taille du ménage et le carré de la taille du ménage servaient de variables continues. Les vingt-sept variables ont été utilisées dans le programme de Huang pour produire les poids de régression. On a fixé à 0.9 la valeur du paramètre M du programme de production de poids et on a arrondi les poids à l'entier naturel le plus près, ces poids étant exprimés en milliers. La somme des poids est de 88,942, chiffre qui correspond au nombre de ménages (en milliers) dans la population. Le poids moyen est de 19,787, le poids le moins élevé, de 6, et le poids le plus élevé, de 47. Le poids le plus élevé équivalait donc à 2.38 fois le poids moyen. La somme des carrés des poids est de 2,317,930. Le produit du carré du poids moyen par la taille de l'échantillon est de 1,759,884. Par conséquent, s'il y a absence de corrélation entre une variable quelconque et les 27 variables, la variance d'estimation calculée à l'aide de poids sera environ 1.32 fois plus élevée que la variance d'un estimateur non pondéré ordinaire.

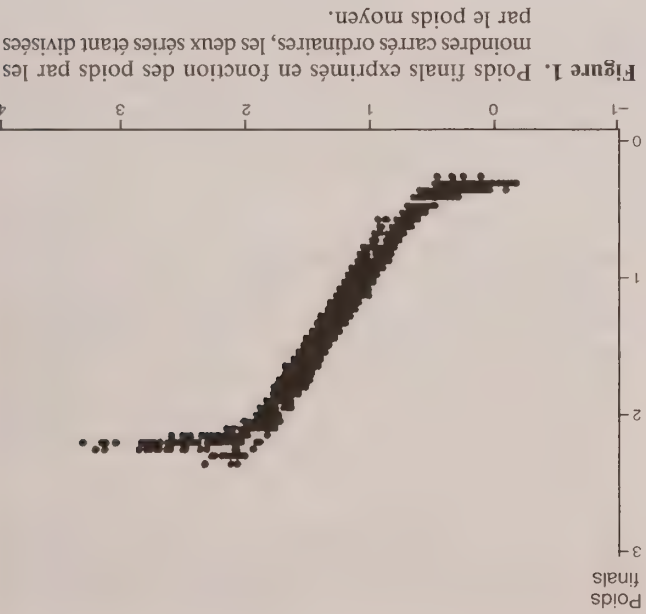


Figure 1. Poids finals exprimés en fonction des poids par les moindres carrés ordinaires, les deux séries étant divisées par le poids moyen.

La figure 1 contient un graphique où sont représentés, sur un axe, les poids de régression calculés au moyen de l'algorithme de Huang et, sur l'autre axe, les poids par les moindres carrés ordinaires. Comme il y a 4,495 ménages, beaucoup de points ne sont pas apparents. Les deux séries de poids sont normalisées, ayant été divisées par le poids moyen. La moyenne pour des poids ayant subi cette transformation est donc égale à un. Comme 27 variables de contrôle sont utilisées, les poids calculés par l'algorithme de Huang forment une nuée de points le long d'une courbe en S tracée en fonction des poids initiaux. S'il n'y avait qu'une variable de contrôle, les points formeraient une courbe en S au sens strict. Les poids initiaux se rattachant à des observations à la gauche du zéro étaient négatifs. Pour comparer les estimations calculées à l'aide de poids avec les estimations non pondérées, nous utilisons les variables suivantes:

$$\begin{aligned} Y_1 &= \text{nombre total redressé de repas pris à l'extérieur du foyer (repas à l'extérieur);} \\ Y_2 &= \text{valeur monétaire totale des aliments consommés à la maison (aliments à la maison);} \\ Y_3 &= \text{taille du ménage en nombre de personnes équivalent à 21 repas (repas-personnes);} \\ Y_4 &= \text{indicateur pour désigner les ménages qui tiennent la maison (tenue de la maison).} \end{aligned}$$

La taille du ménage en nombre de personnes équivalent à 21 repas est le nombre total redressé de repas préparés avec les réserves alimentaires du ménage dans les 7 derniers jours, divisé par 21. La variable "repas-personnes" est la somme de deux termes: le premier est la somme des portions des repas pris à la maison par chaque membre du ménage durant la semaine d'interview; le second, le nombre de repas servis à des invités, à des pensionnaires ou à des employés durant la semaine d'interview, divisé par 21. Autrement dit,

de mars 1987, voir Bureau of the Census (1987). Les chiffres de population pour les catégories du degré d'urbanisation ont été fournis par l'entrepreneur. Dans notre analyse, nous considérons les chiffres de population estimés comme des chiffres officiels.

Tableau 1
Caractéristiques des ménages dans l'échantillon et la population

Caractéristique	Catégorie	Effectif	Proportion dans l'échantillon	Proportion de la population
-----------------	-----------	----------	-------------------------------	-----------------------------

Période d'interview	Printemps	1,828	40.7	25.0
	Été	678	15.1	25.0
Région	Autome	16.0	717	25.0
	Hiver	1,272	28.3	25.0
Degré d'urbanisation	Nord-Est	905	20.1	21.2
	Midwest	1,172	26.1	24.7
Noyau urbain	Sud	1,567	34.9	34.4
	Ouest	851	18.9	19.6
Banlieue	Noyau urbain	1,064	23.7	31.2
	Région non métropolitaine	2,122	47.2	46.0
Revenu du ménage	< 131%	1,041	23.2	20.0
	131-300%	1,564	34.8	32.2
Coupons alimentaires	en % du seuil de pauvreté	1,108	24.6	25.9
	> 500%	782	17.4	21.8
Propriétaire du logement	Oui	314	7.0	7.4
	Non	4,181	93.0	92.6
Origine raciale du chef du ménage	Oui	2,998	66.7	64.1
	Non	1,497	33.3	35.9
Âge du chef du ménage	Noir	519	11.5	11.1
	Non noir	3,976	88.5	88.9
Chef du ménage	< 25	338	7.5	7.9
	25-39	1,588	35.3	36.1
Exacemement un adulte dans le ménage	40-59	1,369	30.5	30.5
	60-69	660	14.7	13.0
Exacemement deux adultes dans le ménage	70 +	540	12.0	12.6
	Homme et femme	3,057	68.0	60.8
Femme chef de ménage a travaillé	Femme seulement	1,044	23.2	26.0
	Homme seulement	394	8.8	13.2
Présence d'enfants de 7 à 17 ans	Oui	1,792	39.9	41.5
	Non	2,703	60.1	58.5
Présence d'enfants de moins de 7 ans	Oui	1,211	26.9	29.7
	Non	3,284	73.1	70.3
Taille du ménage	Oui	2,616	58.2	54.2
	Non	1,879	41.8	45.8
Taille du ménage	Oui	1,009	22.4	20.1
	Non	3,486	77.6	79.9
Taille du ménage au carré	Oui	1,309	29.1	26.5
	Non	3,186	70.9	73.5
Taille du ménage	(Moyenne)	2,731		2,642
	(Moyenne)	9,546		9,125

Le tableau 1 présente les caractéristiques de la population et de l'échantillon de ménages. L'échantillon initial montrait un déséquilibre en ce qui concerne la période d'interview: près de 41% des interviews avaient eu lieu au trimestre de

Alors, l'expression (2.8), où l'on a substitué w_{gt} à w_{it} , est

Nous nous servons de l'estimateur (2.8) dans nos analyses empiriques.

La formule (2.8) définit les deux effets de l'estimation par régression pour la variance de l'estimation de la

par l'effet de corrélation réduit cette variance tandis que l'accroissement de la somme des carrés des poids l'augmente. Pour comprendre ces effets, considérons un échantillon aléatoire simple. Si la variable y est corrélée avec x , la corrélation tend à réduire la variance de l'estimateur par régression par rapport à celle de l'estimateur simple car

$$E\{(y_i - x_i\beta)^2\} \leq E\{(y_i - E(y_i))^2\}.$$

Si la moyenne empirique des variables de contrôle diffère de la moyenne de ces variables pour la population, alors

$$\sum_{i=1}^n w_i^2 > n^{-1},$$

où n^{-1} est la somme des carrés des poids pour un échantillon aléatoire simple.

Lorsque l'on compare la variance de la moyenne de l'échantillon avec celle de l'estimateur par régression, il ne faut pas oublier que le recours à l'estimation par régression pour des échantillons affectés par la non-réponse a pour but notamment de produire un estimateur moins biaisé que l'estimateur direct. C'est pourquoi, dans la section suivante, nous comparons un estimateur de l'erreur quadratique moyenne de l'estimateur simple avec un estimateur de la variance de l'estimateur par régression.

4. APPLICATION À LA NATIONWIDE FOOD CONSUMPTION SURVEY

La Nationwide Food Consumption Survey de 1987-1988 a été effectuée par le Human Nutrition Information Service du Département de l'agriculture des E.-U. L'échantillon initial était un échantillon stratifié autopondéré d'unités primaires d'échantillonnage (u.p.é.) spatiales réparties dans les 48 États contigus. Les u.p.é. étaient divisées en unités secondaires appelées "segments aréolaires". Les ménages des segments aréolaires participaient à une interview en personne. Les opérations sur le terrain étaient confiées à un entrepreneur lié par contrat au Human Nutrition Information Service et se sont déroulées entre avril 1987 et août 1988.

Environ 37% des unités de logement reconnues occupées ont fourni toute l'information voulue sur la consommation d'aliments au sein du ménage. La taille effective de l'échantillon était de 4,495 ménages. Vu la faible taux de réponse, le Human Nutrition Information Service a décidé de recourir à la pondération par régression pour l'estimation. L'organisme a estimé les chiffres de population pour toutes les caractéristiques sauf le degré d'urbanisation en se fondant sur les données de la Current Population Survey

Ensuite,

$$Y = X\gamma + A, \quad (3.5)$$

où $A = N^{-1} \sum_{i=1}^N a_i$ et $a_i = y_i - x_i\gamma$. Donc, l'estimateur par régression (2.1) sera un estimateur convergent de \bar{Y} si $\text{plim}_{N \rightarrow \infty} A = 0$. La probabilité limite de A sera nulle si la population finie est un échantillon aléatoire tiré d'une population infinie pour laquelle le modèle linéaire

$$y_i = x_i\beta + e_i, \quad E\{e_i\} = 0$$

est valide pour tous les i .

La moyenne A est nulle lorsque $\pi_i^* = \pi_i$ pour tous les i et qu'un élément de x_i est égal à un pour tous les i car dans ces conditions

$$\gamma = \beta = \left(\sum_{i=1}^N x_i' x_i \right)^{-1} \sum_{i=1}^N x_i' y_i \quad (3.6)$$

et $\sum_{i=1}^N (y_i - x_i\beta) = 0$. Une condition suffisante pour que A soit nulle est l'existence d'un vecteur ligne c qui soit

$$cx_i' = \pi_i^{*-1} \pi_i = p_i^{-1}, \quad (3.7)$$

pour $i = 1, 2, \dots, N$. Donc, si le rapport de la probabilité des variables de contrôle, l'estimateur par régression est nominal à la probabilité vraie est une fonction linéaire des variables de contrôle, l'estimateur par régression est s'appliquant à des suites comme le définit Fuller (1975). Pour que l'équation (3.7) soit satisfaite, il faudrait, par exemple, que les éléments de x_i soient des variables fictives qui définissent des sous-groupes et que les probabilités de réponse soient les mêmes à l'intérieur de chaque sous-groupe. On décrit parfois cette situation en disant que des éléments sont absents de façon aléatoire dans chaque sous-groupe. Dans l'analyse empirique, nous nous servons de l'hypothèse que $A = 0$ comme hypothèse de travail.

Dans un problème de régression, il y a certaines hypothèses qui ne peuvent être vérifiées à l'aide des données de l'échantillon. Dans la régression par les moindres carrés ordinaires par exemple, les résidus $\hat{e}_i = y_i - x_i'\hat{\beta}$ ne sont pas corrélés avec x_i et ne peuvent donc pas servir à vérifier l'hypothèse selon laquelle les erreurs vraies sont non corrélées avec x . Par conséquent, dans une enquête avec non-réponse, on doit trouver des variables de contrôle qui sont corrélées avec y ou que l'on croit être corrélées avec les probabilités de réponse. Toutefois, un ensemble partiel de variables de contrôle ne suffit pas pour affirmer que l'estimation par régression a contribué à supprimer le biais.

Dans la pratique, il est souvent possible de trouver des variables x qui sont corrélées avec la probabilité de réponse ou avec les variables y . Par exemple, dans la Nationwide Food Consumption Survey de 1987-1988, on a enregistré un faible taux de réponse chez les ménages à revenu élevé. Par conséquent, l'utilisation de variables ayant trait au revenu du ménage dans un estimateur par régression est

Supposons aussi que

$$\text{plim}_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N (y_i - x_i'\gamma_g) = 0.$$

où

$$\hat{\beta}_g = \left[\sum_{i=1}^N x_i' \pi_i g_i x_i \right]^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} g_i y_i,$$

où les g_i sont fonction des x_i . Supposons que

$$\text{plim}_{N \rightarrow \infty} \hat{\beta}_g = \gamma_g,$$

où les g_i sont fonction des x_i . Supposons que

$$w_{gi} = X \left[\sum_{i=1}^N x_i' \pi_i^{-1} g_i x_i \right]^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} g_i,$$

exemple, définissons les poids par l'expression

L'estimateur (2.8) peut aussi convenir lorsque les poids de régression sont calculés autrement que par (2.4). Par

comme estimateur de la variance de l'estimation de la moyenne de y .

Si l'estimateur par régression en estimant la variance de $\sum_{i=1}^N x_i' \pi_i^{-1} a_i$. Si nous supposons que les probabilités conditionnelles de réponse relatives à une unité primaire d'échantillonnage sont indépendantes de celles relatives à toutes les autres et qu'au moins une unité est observée dans chaque u.p.é. échantillonnée, alors l'estimateur (2.8) est toujours indiqué comme estimateur de la variance de l'estimation de la

$$G = T^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} x_i$$

et

$$T = \sum_{i=1}^N \pi_i^{-1}$$

Suivant des hypothèses raisonnables,

$$G = T^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* x_i.$$

et

$$T = \sum_{i=1}^N \pi_i^{-1} \pi_i^*.$$

où a_i est défini en (3.5),

$$\hat{\beta} - \gamma = G^{-1} T^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} a_i,$$

être définie approximativement par la formule

des caractéristiques qui sont corrélées avec le revenu. L'erreur contenue dans $\hat{\beta}$, comme estimateur de γ , peut

censée réduire le biais de l'estimation de la moyenne pour

ont décrit la méthode de production de poids et montré que la distribution de l'estimateur modifiée pour de grands échantillons est identique à celle de l'estimateur par régression ordinaire. Voir aussi Goebel (1976).

L'algorithme de Huang (1978) est une méthode itérative fondée sur le principe des moindres carrés généralisés. Cet algorithme a pour but de produire un ensemble de poids w_i , $i = 1, 2, \dots, n$, qui satisfont (2.5) et qui diffèrent peu des poids initiaux, l'écart étant une fonction du poids initial. L'algorithme de Huang vise à calculer des poids w_i qui satisfont l'inéquation

$$(1 - M) \max_{1 \leq i \leq n} w_i \pi_i^{-1} \leq (1 + M) \min_{1 \leq i \leq n} w_i \pi_i^{-1},$$

où le paramètre M , $0 < M \leq 1$, est défini par l'utilisation de poids. Ceux-ci sont des poids de régression pondérés, où chaque observation est affectée d'un facteur de contrôle: facteur peu élevé pour les observations qui ont un poids primaire faible ou élevé; facteur relatif grand pour les observations qui ont un poids primaire voisin de π_i^{-1} . Les poids de la deuxième série sont soumis à une vérification et, s'ils ne satisfont pas aux conditions, on modifie les facteurs de contrôle, et ainsi de suite. L'algorithme est décrit en annexe.

La pondération de contrôle utilisée dans l'algorithme de Huang a beaucoup en commun avec les méthodes de réduction d'influence et les méthodes de régression robustes, c'est-à-dire que dans l'estimateur final, la pondération de contrôle a pour effet de réduire la part des observations éloignées de la moyenne dans l'estimation du vecteur-pente. Voir Hampel (1978), Krasker (1980) et Mallows (1983). Parmi les ouvrages récents portant sur ce type d'estimateurs pour des échantillons d'enquête, notons Deville et Särndal (1992), Akkerboom, Sikkink et van Herk (1991), Hülliger (1993) et Singh (1993).

Il n'est pas toujours possible de créer des poids qui en même temps répondent aux critères et satisfont l'équation (2.5). Par exemple, si toutes les observations de poids excèdent la moyenne, il n'y aura aucun ensemble de poids positifs dont la somme sera égale à un et qui satisfont aussi $\sum_{i=1}^n w_i = X_2$. Le programme de production de poids s'interromptra donc si, après un nombre déterminé d'itérations, il n'est pas possible d'obtenir des poids qui répondent aux critères établis.

Dans certains cas, il est souhaitable de limiter les poids à des entiers non négatifs, par exemple lorsqu'on établit des estimations de totaux et que la population renferme des unités bien définies, comme des personnes. L'utilisation de poids entiers non négatifs donne alors des estimations plus satisfaisantes en ce sens qu'elles sont palpables. On peut créer des poids en nombre entier de sorte qu'il ne soit pas nécessaire d'effectuer un arrondissement lorsqu'on construit des tableaux. Avec de tels poids, tous les tableaux à multiples entrées seront automatiquement cohérents.

Le programme de Huang contient une fonction qui permet d'arrondir les poids en nombre réel sans en modifier

la somme. Après l'arrondissement, l'équation (2.5) ne sera en général plus parfaitement exacte. Nous avons observé qu'en itérant l'algorithme de Huang en se servant des poids primaires en nombre entier comme poids initiaux, on pouvait obtenir des poids en nombre entier qui satisfaisaient à peu de choses près l'équation (2.5). La question de l'arrondissement est traitée dans Cox (1987), Cox et Ernst (1982) et Fagan, Greenberg et Hemmig (1988).

3. ESTIMATION PAR RÉGRESSION EN SITUATION DE NON-RÉPONSE

À l'origine, la théorie de l'estimation par régression supposait que l'échantillon étudié était un échantillon probabiliste de la population. Or, on reconnaît depuis longtemps déjà que l'estimation par régression peut servir à réduire le biais qui découle des faiblesses de la méthode de collecte des données. La plus notable de ces faiblesses est la non-réponse. Dans tous les grands échantillons constitués de personnes, il y a des individus qui ne donnent pas l'information voulue. Si les non-répondants n'ont pas les mêmes caractéristiques que les répondants, les estimations directes établies à l'aide des données recueillies seront biaisées. Si l'on possède de l'information supplémentaire, l'estimation par régression est une méthode pour réduire le biais. La grandeur de la réduction dépendra du rapport entre les variables de contrôle, les variables étudiées et les probabilités de réponse. Voir Little et Rubin (1987) pour une analyse générale de ces questions.

Posons π_i^* comme la probabilité conjointe de sélection et de réponse, exprimée comme le produit de π_i par la probabilité conditionnelle d'observer l'unité étant donné que celle-ci a été échantillonnée. Alors,

$$E \left\{ \sum_{i=1}^n x_i' \pi_i^{-1} x_i \mid \xi_N \right\} = \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* x_i \quad (3.1)$$

et

$$E \left\{ \sum_{i=1}^n x_i' \pi_i^{-1} y_i \mid \xi_N \right\} = \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* y_i \quad (3.2)$$

où l'espérance dépend de la population finie donnée ξ_N et n est la taille effective de l'échantillon. En situation de non-réponse, le ratio $p_i = \pi_i^* \pi_i^{-1}$ est la probabilité de réponse pour l'individu i . Par conséquent, dans des conditions comme celles définies par Fuller (1975),

$$\text{plim}(\hat{\beta} - \gamma) = 0, \quad n \rightarrow \infty \quad (3.3)$$

où $\hat{\beta}$ est défini en (2.2) et

$$\gamma = \left(\sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* x_i \right)^{-1} \sum_{i=1}^N x_i' \pi_i^{-1} \pi_i^* y_i. \quad (3.4)$$

cherche à obtenir un estimateur de la moyenne de y . Nous supposons que le premier élément de x_i est égal à un pour tous les i . Par conséquent, le premier élément de X est aussi égal à un. Le vecteur $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ est parfois appelé vecteur des variables de contrôle. Un estimateur par régression de la moyenne de y est

$$(2.1) \quad \bar{y}_r = \bar{X}\hat{\beta},$$

où

$$(2.2) \quad \hat{\beta} = \left(\sum_{i=1}^n x_i' \pi_i^{-1} x_i \right)^{-1} \sum_{i=1}^n x_i' \pi_i^{-1} y_i,$$

$\sum x_i' \pi_i^{-1} x_i$ étant supposée régulière. Cette définition de l'estimateur par régression est conforme à celle donnée par Mickey (1959), qui proposait de limiter l'usage du terme "estimateur par régression" aux estimateurs qui sont invariants en position et en échelle. L'estimateur (2.1) peut aussi s'écrire sous la forme

$$(2.3) \quad \bar{y}_r = \sum_{i=1}^n w_i y_i,$$

où

$$(2.4) \quad w_i = X \left(\sum_{i=1}^n x_i' \pi_i^{-1} x_i \right)^{-1} x_i' \pi_i^{-1},$$

les poids ayant la propriété suivante:

$$(2.5) \quad \sum_{i=1}^n w_i x_i = \bar{X}.$$

Les poids définis en (2.4) sont assez faciles à calculer et, une fois qu'ils l'ont été, ils peuvent servir à l'estimation de n'importe quelle variable y . Si on remplace le vecteur x_j par le vecteur

$$(1, z_j) = (1, x_{j2} - \bar{X}_{x_{j2}}, x_{j3} - \bar{X}_{x_{j3}}, \dots, x_{jk} - \bar{X}_{x_{jk}}), \quad (2.6)$$

on peut exprimer l'estimateur sous la forme

$$(2.7) \quad \bar{y}_r = \bar{y}_\pi + (\bar{Z} - \bar{z}_\pi) \hat{\beta}_z = \bar{y}_\pi - \bar{z}_\pi \hat{\beta},$$

où $\bar{Z} = 0$ est la moyenne de population de z_j , $\bar{z}_\pi = \bar{x}_\pi - \bar{X}$,

$$(\bar{y}_\pi, \bar{z}_\pi) = \left(\sum_{i=1}^n \pi_i^{-1} \right)^{-1} \sum_{i=1}^n \pi_i^{-1} (y_i, z_i)$$

et

$$\hat{\beta}_z = \left[\sum_{i=1}^n (z_i - \bar{z}_\pi)' \pi_i^{-1} (z_i - \bar{z}_\pi) \right]^{-1} \sum_{i=1}^n (z_i - \bar{z}_\pi)' \pi_i^{-1} y_i.$$

Dans l'expression (2.7), \bar{y}_r représente l'ordonnée à l'origine dans la régression de y par rapport à z . Par conséquent, la théorie exposée par Fuller (1975) sur les coefficients de régression s'applique à l'estimateur par régression de la moyenne. Si l'effectif de la population est connu et désigné par N , le total estimé pour la population est $N\bar{y}_r$. En définissant une série de populations et d'échantillons, on peut montrer que l'estimateur (2.1) est un estimateur convergent de la moyenne de y . Voir, par exemple, Fuller (1975). L'estimateur de la variance de l'estimateur par régression est une fonction des probabilités conjointes. Considérons un échantillon à deux degrés stratifié et remplaçons l'indice inférieur i par l'indice triple ℓ/t . Alors, en omettant le facteur de correction pour population finie, nous avons l'estimateur de variance suivant:

$$V\{\bar{y}_r\} = (n - k)^{-1} \sum_{\ell=1}^L n_{\ell}^{-1} (n_{\ell} - 1)^{-1} n_{\ell} \sum_{t=1}^{n_{\ell}} (d_{\ell t} - d_{\ell..})^2, \quad (2.8)$$

où

$$d_{\ell t} = \sum_{m_{\ell t}} w_{\ell t} (y_{\ell t} - x_{\ell t} \hat{\beta}),$$

$$d_{\ell..} = n_{\ell}^{-1} \sum_{j=1}^{n_{\ell}} d_{\ell j},$$

n_{ℓ} étant le nombre d'unités primaires-échantillon dans la strate ℓ , $m_{\ell t}$, le nombre d'éléments de l'échantillon dans l'unité primaire d'échantillonnage j de la strate ℓ , $\hat{\beta}$, le vecteur de coefficients défini en (2.2), n , l'effectif de l'échantillon et $w_{\ell t}$, le poids de l'élément t dans l'unité primaire d'échantillonnage j de la strate ℓ . Le facteur $n - k$ est utilisé par analogie avec le diviseur contenu dans l'estimateur sans biais de la variance de l'erreur pour la régression ordinaire. Lorsque le vecteur des variables de contrôle est exprimé sous la forme définie en (2.6), l'équation (2.8) correspond à la variance estimée du premier élément de $\hat{\beta}$, l'ordonnée à l'origine estimée. L'estimateur (2.8) avait été proposé dans Hidiroglou, Fuller et Hickman (1976) et sa convergence, vérifiée dans Fuller (1975). Voir aussi Särndal, Swensson et Wretman (1989). Les estimateurs construits avec les poids (2.4) conviennent bien pour de grands échantillons, mais sont peut-être moins intéressants avec de petits échantillons. Comme les poids sont des fonctions linéaires de x_i , il se peut que certains d'entre eux soient négatifs. Or, des poids négatifs peuvent faire que des estimations de paramètres positifs seront négatives. Husain (1969) a exploré des méthodes pour construire des poids de régression non négatifs. Huang (1978), pour sa part, a élaboré un programme informatique pour produire de tels poids. Huang et Fuller (1978)

Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988

WAYNE A. FULLER, MARIE M. LOUGHIN et HAROLD D. BAKER¹

RÉSUMÉ

Les auteurs appliquent une méthode de production de poids de régression à la Nationwide Food Consumption Survey réalisée en 1987-1988 par le Département de l'agriculture des États-Unis. Ils ont eu recours à l'estimation par régression des moindres carrés généralisés, modifiés de telle manière qu'ils soient tous positifs et que les plus élevés d'entre eux soient plus petits que les poids établis par les moindres carrés. Les auteurs montrent qu'en situation de non-réponse l'estimateur par régression peut avoir une erreur quadratique moyenne beaucoup moins élevée que l'estimateur direct ordinaire.

MOTS CLÉS: Poids non négatifs; cohérence.

1. INTRODUCTION

Souvent, quand on procède à un échantillonnage, on

la population mais on ne se sert pas, pour le prélèvement de l'échantillon, des valeurs de ces variables pour les unités de cette population. Même si cette information n'est pas utilisée dans le plan d'échantillonnage, il est très souhaitable d'intégrer les moyennes de population dans la procédure d'estimation. Les méthodes d'estimation courantes qui utilisent de l'information supplémentaire sont l'estimation par quotient, la stratification a posteriori, l'estimation par régression et l'estimation par la méthode itérative du quotient. L'estimation par régression est la méthode la plus générale en ce sens qu'elle peut traiter des variables auxiliaires multidimensionnelles, continues ou discrètes. La stratification a posteriori peut être considérée comme un cas particulier de l'estimation par régression, où les variables de régression sont des variables indicatrices pour les strates formées a posteriori. La méthode itérative du quotient, aussi appelée ajustement proportionnel itératif, est limitée à de l'information supplémentaire sous forme de catégories discrètes. Sur cette méthode itérative du quotient, on pourra lire Deming et Stephan (1940), Stephan (1942), El-Badry et Stephan (1955), Ireland et Kullback (1968), Darroch et Ratcliff (1972), Brackstone et Rao (1979) et Oh et Scheuren (1987).

Watson (1937), Cochran (1942) et Jessen (1942) sont les premiers ouvrages où l'on traite d'estimation par régression. La théorie élémentaire de cette forme d'estimation se trouve dans Cochran (1977, chap. 7). De nombreux auteurs se sont intéressés à l'estimation par régression appliquée à des échantillons d'enquête: Mickey (1959), Fuller (1975), Royall et Cumberland (1981), Isaki et Fuller (1982), Wright (1983), Luey (1986), Alexander (1987), Bethlehem et Keller (1987), Copeland, Peitzmeier et Hoy

Dans l'application décrite plus bas, la non-réponse nous amène à recourir à l'estimation par régression, et c'est cette expérience que nous décrivons dans la section 3. Dans la section 2, nous présentons l'estimateur par régression ordinaire ainsi qu'une version modifiée de celui-ci qui produit des poids positifs. Enfin, dans la section 4, nous appliquons la méthode de pondération à la Nationwide Food Consumption Survey.

2. ESTIMATEUR PAR RÉGRESSION

Afin de présenter l'estimateur par régression que nous allons utiliser dans cette étude, nous supposons qu'un échantillon de n unités est prélevé et que la probabilité de sélection pour l'unité i est π_i . Pour les besoins de notre exposé, nous posons comme condition suffisante que π_i soit proportionnelle aux probabilités de sélection. Il peut s'agir d'un échantillon stratifié à deux degrés et l'unité peut être soit l'unité primaire d'échantillonnage, soit l'unité d'observation. Dans notre application, il s'agit de l'unité d'observation. Supposons qu'un vecteur de moyennes de population de dimension k , désigné par $\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$, est connu, qu'un vecteur $(y_i, x_{i1}, x_{i2}, \dots, x_{ik})$ est observé pour chaque unité de l'échantillon et qu'on

¹ Wayne A. Fuller, Marie M. Loughin et Harold D. Baker, Iowa State University.

BIBLIOGRAPHIE

- BRYANT, E.C., HARTLEY, H.O., et JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association*, 55, 105-124.
- CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.
- GOODMAN, R., et KISH, L. (1950). Controlled selection-a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- JESSEN, R.J. (1970). Probability sampling with marginal constraints. *Journal of the American Statistical Association*, 65, 776-795.
- RAO, J.N.K., et NIGAM, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.
- RAO, J.N.K., et NIGAM, A.K. (1992). 'Optimal' controlled sampling: a unified approach. *Revue Internationale de Statistique*, 60, 89-98.
- WATERTON, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.

$$\frac{1}{2n^2D} \sum_{\{c:n_c(s) \geq 2\}} \left(\frac{n^2}{N^2} - A_c \right) \sum_{n_c(s)} \sum_{k=1}^k \sum_{k'=1}^{k'} 2N_c^2$$

$$\frac{(y^{ck} - y^{ck'})^2}{2n_c(s) \{n_c(s) - 1\}}$$

où $D =$ le nombre de cellules c telles que $n_c(s) \geq 2$.

Le raisonnement ci-dessus exige que $B_{cc'} > 0$. Si

$$I_c = I_{c'} = 0,$$

alors, d'après (5.1), il faut que

$$r_{cc'} = \sum n_c(s) n_{c'}(s) p(s) > 0, \quad (5.2)$$

qui est linéaire en $p(s)$. La contrainte (5.2) peut être traitée par programmation linéaire, si désire. Une telle contrainte existera pour chaque paire c, c' telle que $I_c = I_{c'} = 0$.

6. CONCLUSION

Nous avons proposé une méthode fondée sur la programmation linéaire pour le traitement de stratifications multidimensionnelles, en nous appuyant sur les idées de Rao et Nigam (1990, 1992). Il s'agit d'une méthode simple, qui offre beaucoup de souplesse en permettant de choisir toute une gamme d'objectifs différents par l'entremise de la fonction de perte $w(s)$, et en permettant tout un éventail de contraintes, comme le fait d'exiger que les probabilités de sélection conjointes de toutes les cellules soient positives. Le principal obstacle pratique à l'application de la méthode est que son traitement informatique peut rapidement devenir coûteux, voire impossible, à mesure qu'augmente le nombre de cellules de la stratification multidimensionnelle. Certaines suggestions ont été faites en vue de réduire l'ampleur des calculs. Il serait utile que d'autres recherches soient entreprises dans ce domaine. Pour les cas où les exigences du traitement informatique seraient prohibitives, la méthode de Causey et coll. (1985) demeure une solution de rechange.

REMERCIEMENTS

Les auteurs tiennent à remercier Wesley Yung pour son aide touchant la programmation informatique, ainsi que J.N.K. Rao pour ses précieux commentaires aux premiers stades du travail. R.R. Sitter est subventionné par le Conseil de recherches en sciences naturelles et en génie du Canada. Nous exprimons nos remerciements aux arbitres et au rédacteur associé pour leurs suggestions et leurs commentaires utiles.

ANNEXE

Preuve de (4.7) pour la méthode proposée

Notons que

$$\text{Cov}(n_{ij}(s), n_{i'j'}(s)) = E(n_{ij}(s) n_{i'j'}(s))$$

$$- E(n_{ij}(s)) E(n_{i'j'}(s)).$$

L'équation (2.1) indique que $E(n_{ij}(s)) = n_j P_j$. Par définition

$$E(n_{ij}(s) n_{i'j'}(s)) = \sum_s n_{ij}(s) n_{i'j'}(s) p(s).$$

Donc

$$\sum_j E(n_{ij}(s)) E(n_{i'j'}(s)) = n_i^2 P_{i'j'} \sum_j P_j = n_i^2 P_{i'j'} P_i, \quad (7.1)$$

et

$$\sum_j E(n_{ij}(s) n_{i'j'}(s)) = \sum_s \sum_j n_{ij}(s) n_{i'j'}(s) p(s)$$

$$= \sum_s p(s) n_{i'j'}(s) \sum_j n_{ij}(s). \quad (7.2)$$

Supposons que la solution du problème d'optimisation linéaire (2.2) soit égale à zéro, avec $w(s)$ donné par (2.9). Dans ce cas, $\sum_j n_{ij}(s) = n_i$, $(s) = n_j$, et il résulte de (7.2) que

$$\sum_j E(n_{ij}(s) n_{i'j'}(s)) = \sum_s p(s) n_{i'j'}(s) n_j P_i$$

$$= n_j P_i \sum_s n_{i'j'}(s) p(s)$$

$$= n_j P_i E(n_{i'j'}(s)) = n_j P_i P_{i'j'}.$$

(7.3)

Les équations (7.1) et (7.3), ensemble, font en sorte que $\sum_j \text{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0$. On peut montrer de la même manière que

$$\sum_i \text{Cov}(n_{ij}(s), n_{i'j'}(s)) = \sum_{i'} \text{Cov}(n_{ij}(s), n_{i'j'}(s))$$

$$= \sum_{j'} \text{Cov}(n_{ij}(s), n_{i'j'}(s)) = 0.$$

Les probabilités de sélection conjointe de paires d'unités de cellules différentes c et c' sont

$$\pi_{ck,ck'} = E[n_c(s)n_{c'}(s)/(N_cN_{c'})]$$

$$= [I_c I_{c'} + r_{c'} I_c + r_c I_{c'} + r_{cc'}]/(N_c N_{c'}) = B_{cc'},$$

(5.1)

par exemple, où $r_{cc'} = E[\bar{n}_c(s)\bar{n}_{c'}(s)]$.

Par conséquent, un estimateur non biaisé de $V(\bar{y}(s))$, de la forme Sen-Yates-Grundy, peut être construit de la manière habituelle.

En pratique, toutefois, nous voulons tenir compte de situations où $nP_c \leq 1$ pour certaines cellules c . Dans ce cas, une hypothèse que nous pourrions faire, comme l'ont fait Bryant et coll. (1960, section 7), afin de trouver un estimateur de la variance, c'est de considérer que la variance de X pour l'ensemble de la population dans chaque cellule c est une valeur constante, disons S_c^2 .

Obtenons d'abord la variance de $\bar{y}(s)$ dans le cas général

$$V(\bar{y}(s)) = \frac{1}{2n^2} \sum_c \sum_{k \neq k'} \left(\frac{N^2}{n^2} - A_c \right) (y_{ck} - y_{ck'})^2 + \frac{1}{2n^2} \sum_{c \neq c'} \sum_{k,k'} \left(\frac{N^2}{n^2} - B_{cc'} \right) (y_{ck} - y_{c'k'})^2.$$

Pourvu que $B_{cc'} > 0 \forall c, c'$, nous pouvons estimer sans biais le second terme sous la forme

$$\frac{1}{2n^2} \sum_c \sum_{k=1}^A \sum_{k'=1}^A \left(\frac{N^2}{n^2} - B_{cc'} \right) (y_{ck} - y_{c'k'})^2,$$

où $A = \{c, c' : n_c(s) \geq 1, n_{c'}(s) \geq 1, c \neq c'\}$.

Le premier terme peut s'écrire ainsi

$$\frac{1}{2n^2} \sum_c \left(\frac{N^2}{n^2} - A_c \right) 2N_c^2 S_c^2.$$

Pour tout c tel que $n_c(s) \geq 2$

$$E \left(\sum_{n_c(s)} \sum_{\substack{k=1 \\ k' \neq k}}^{k=1} \frac{2n_c(s) \{n_c(s) - 1\}}{(y_{ck} - y_{ck'})^2} \middle| n_c(s) \right) = S_c^2.$$

Donc, pourvu qu'au moins un $n_c(s)$ soit ≥ 2 , un estimateur non biaisé du premier terme est donné par

Le fait que v_2 , dans le plan de Bryant et coll., s'accroisse rapidement avec l'augmentation de $|\gamma|$ semble attribuable à la plus grande variabilité de chaque n_{ij} dans ce plan. Notons qu'il aurait été possible de réduire quelque peu cette variabilité en appliquant plutôt au tableau 2 une variante de la méthode de Bryant et coll., comme nous l'avons fait pour la méthode proposée, mais il aurait fallu calculer des poids G_{ij} rajustés pour \bar{y}_U , et il serait difficile de traiter les valeurs 0.0 des cellules du tableau 2. Toutefois, même après avoir fait cela, les n_{ij} pourraient toujours, dans ce plan, prendre des valeurs autres que 0 et 1; par exemple, m_2 pourrait prendre les valeurs 0, 1 ou 2. Cette variabilité supérieure des n_c est inhérente à la méthode de Bryant et coll. Par exemple, supposons que $n_1 = n_2 = 5$. Selon la méthode de Bryant et coll., n_{11} peut prendre les valeurs 0, 1, 2, 3, 4 ou 5, tandis que selon la méthode proposée, il ne peut prendre que les valeurs $[n_{11}]$ ou $[n_{11}] + 1$. Si $n_{11} < 1$, la technique utilisée pour passer du tableau 1 au tableau 2 n'améliorera pas les choses.

5. ESTIMATION DE LA VARIANCE

Dans la présente section, nous allons nous pencher sur l'estimation de la variance pour la méthode que nous proposons. En utilisant (4.1) et en nous rappelant de la contrainte (2.1), nous pouvons évidemment écrire

$$\pi_{ck} = E[n_c(s)/N_c] = n/N.$$

La probabilité de sélection conjointe de deux unités, k et k' , de la même cellule c est

$$\pi_{ck,ck'} = E[n_c(s)\{n_c(s) - 1\}/\{N_c(N_c - 1)\}].$$

Supposons que $n_c(s) = I_c + \bar{n}_c(s)$ où I_c est l'entier fixe $[nP_c]$ et $\bar{n}_c(s) = 0$ ou 1.

Si $nP_c \leq 1$, alors $I_c = 0$ et $\pi_{ck,ck'} = 0$. Par conséquent, une condition nécessaire pour qu'une estimation non biaisée de la variance soit possible est que $nP_c > 1$ pour toutes les cellules c . D'autre part, si cette condition est vérifiée, il en résulte que $n_c(s) \geq 1$ pour toutes les cellules c , de sorte que la probabilité de sélection de toute paire d'unités de cellules différentes est aussi toujours positive. Bref, cette condition est nécessaire et suffisante pour qu'une estimation non biaisée de la variance soit possible.

Quand cette condition est vérifiée, nous obtenons

$$\pi_{ck,ck'} = I_c(I_c + 2r_c - 1)/[N_c(N_c - 1)] = A_c,$$

par exemple, où $r_c = E[\bar{n}_c(s)] = nP_c - I_c$.

Si N est grand, cela équivaut approximativement à

$$\pi^{ckc'k'} \doteq \frac{E(n_c P_c)}{E(n_c)} - \frac{N^2 P_c^2}{I^{[c=c']}}, \quad (4.3)$$

où

$$I^{[c=c']} = \begin{cases} 1 & \text{si } c = c', \\ 0 & \text{si } c \neq c', \end{cases}$$

Les espérances de notre plan ne seront pas les mêmes que

celles du plan de Bryant et coll., de sorte que les π^{ck} et les $\pi^{ckc'k'}$ seront différents. En gardant ce fait à l'esprit, nous pouvons obtenir la variance de \bar{y}_j, \bar{y}_U et \bar{y}_B sous une forme généralisée en termes des valeurs π^{ck} et $\pi^{ckc'k'}$, et donc disposer d'une base de comparaison. Pour ce faire, prenons un estimateur de la forme $\bar{z} = \sum_k w_c y_{ck}/n$, où les valeurs w_c sont des constantes connues fixes, indépendantes de k . Si nous nous limitons au cas où $n_j = n_{j'}$ et $n_{j'} = n_{j''}$, c'est-à-dire que les valeurs marginales sont entières, les estimateurs donnés par Bryant et coll. ainsi que notre estimateur sont tous de cette forme. Nous supposons dorénavant que c'est le cas. Après remplacement de l'indice c par \bar{y} pour la stratification bidimensionnelle, \bar{y}_U et \bar{y}_B sont de la même forme que \bar{z} avec $w_c = w_{\bar{y}} = G_{\bar{y}} = P_{\bar{y}}/(P_{\cdot} P_{\cdot j})$ et $w_c = w_{\bar{y}} = 1$ respectivement. L'estimateur \bar{y} est également de la même forme que \bar{z} , avec $w_c = w_{\bar{y}} = 1$.

Nous pouvons maintenant obtenir une forme générale pour la variance de \bar{z} , en nous rappelant que les π^{ck} et les $\pi^{ckc'k'}$ de notre plan ne seront pas les mêmes que celles du plan de Bryant et coll.:

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \sum_{c'} \sum_k \sum_{k'} (\pi^{ck} \pi^{c'k'} - \pi^{ckc'k'})$$

$$(4.4) \quad (w_c y_{ck} - w_{c'} y_{c'k'})^2.$$

En utilisant (4.1) et (4.3), on obtient

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c \sum_{c'} \sum_k \sum_{k'} \frac{N^2 P_c^2}{w_c^2 E(n_c)} (y_{ck} - y_{c'k'})^2$$

$$- \frac{1}{2n^2} \sum_{c'} \sum_c \frac{N^2 P_c P_{c'}}{\text{Cov}(n_c, n_{c'})} \sum_{k'} \sum_{k''}$$

$$(4.5) \quad (w_c y_{ck} - w_{c'} y_{c'k'})^2.$$

Notons que

$$\sum_k \sum_{k'} (y_{ck} - y_{c'k'})^2 = 2N^2 P_c^2 S_c^2$$

et que

$$\sum_{k'} \sum_{k''} (w_c y_{ck} - w_{c'} y_{c'k'})^2 = N^2 P_c P_{c'}$$

$$[w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2 + (w_c Y_{c'} - w_{c'} Y_c)^2],$$

où S_c^2 désigne la variance de la population de la cellule c . L'équation (4.5) se réduit alors à

$$V(\bar{z}) = \frac{1}{2n^2} \sum_c w_c^2 E(n_c) S_c^2 - \frac{1}{2n^2} \sum_{c'} \sum_c \text{Cov}(n_c, n_{c'}) [w_c^2 S_c^2 + w_{c'}^2 S_{c'}^2]$$

$$(4.6) \quad = v_1 + v_2, \quad \text{par exemple.}$$

Le premier terme, v_1 , peut être interprété comme la variance stratifiée habituelle pour les tailles d'échantillon fixes $E(n_c)$ à l'intérieur des "strates" bidimensionnelles (évidemment, dans notre cas, les $E(n_c)$ ne seront pas généralement des entières). Le second terme, v_2 , peut être interprété comme l'accroissement de variance attribuable à la variabilité des n_c et à la corrélation qui existe entre eux. Nous y reviendrons à la fin de la présente section. Nous revenons maintenant à la notation $c = \bar{y}$ et comparons les variances pour la stratification bidimensionnelle. Examinons d'abord v_1 dans (4.6). Pour la méthode de Bryant et coll. $E(n_{\bar{y}}) = n_{P_{\cdot} P_{\cdot j}} = \sum_i \sum_j \sum_k G_{\bar{y}} y_{ijk}/n$, $G_{\bar{y}} = P_{\bar{y}}/(P_{\cdot} P_{\cdot j})$ et $\bar{y}_B = \sum_i \sum_j \sum_k \Delta_k y_{ijk}/n$.

Ainsi,

$$v_1(\bar{y}_U) = \sum_j \sum_i P_{\bar{y}} G_{\bar{y}} S_{\bar{y}}^2/n,$$

(ce qui est identique au premier terme de l'équation (12) de Bryant et coll.) et

$$v_1(\bar{y}_B) = \sum_j \sum_i P_{\cdot j} P_{\cdot} S_{\bar{y}}^2/n.$$

Dans le cas de notre méthode, $E(n_{\bar{y}}) = n_{P_{\cdot j}}$ et $\bar{y} = \sum_i \sum_j \Delta_k y_{ijk}/n$, de sorte que

$$v_1(\bar{y}) = \sum_j \sum_i P_{\bar{y}} S_{\bar{y}}^2/n.$$

Examinons maintenant v_2 . Il n'est pas difficile de montrer qu'aussi bien pour la méthode de Bryant et coll. que pour la nôtre (voir l'annexe)

$$(4.7) \quad \sum_i \text{Cov}(n_{\bar{y}}, n_{i'j'}) = \sum_j \text{Cov}(n_{\bar{y}}, n_{i'j'}) = 0.$$

En tenant compte de ce fait et en remplaçant c et c' par \bar{y} et $i'j'$ respectivement dans l'expression v_2 donnée en (4.6), nous réduisons v_2 à

$$v_2 = \frac{1}{2n^2} \sum_j \sum_{i'} \sum_{j'} \sum_{i''j''} \text{Cov}(n_{\bar{y}}, n_{i'j'}) (w_{\bar{y}} w_{i'j''} - w_{\bar{y}} w_{i'j'})^2.$$

Exemple 2: Jessen (1970)

Jessen (1970) a proposé deux méthodes, pour la stratification bidimensionnelle et la stratification tridimensionnelle. Les deux méthodes, très complexes, obligent à trouver l'ensemble d'échantillons qui correspondent exactement aux marges. Aucune d'elle ne donne nécessairement une solution. Jessen (1970) applique les deux méthodes à un exemple hypothétique simple pour lequel chacune produit une solution. Cet exemple est présenté au tableau 4. Dans ce cas, puisque tous les $nP_{ij} < 1$, les problèmes de programmation linéaire définis respectivement par (2.1), (2.2) et (2.3) et par (2.4), (2.5) et (2.6) sont identiques. Nous avons appliqué notre méthode à ce problème, en utilisant encore la fonction $w(s)$ définie en (2.9). En essayant plusieurs valeurs de départ dans le programme d'optimisation, nous avons pu obtenir trois solutions différentes, chacune rendant (2.2) égal à zéro et satisfaisant aux contraintes. Ces solutions sont présentées au tableau 5. Les deux premières sont les mêmes que celles obtenues respectivement par la méthode 2 et par la méthode 3 de Jessen.

Tableau 4

Exemple 2: Jessen (1970)

Nombres probables d'unités de l'échantillon dans les cellules en vertu d'une stratification proportionnelle avec $n = 6$

Lignes $nP_{i.}$	Colonnes		
	1	2	3
1	0.8	0.5	0.7
2	0.7	0.8	0.5
3	0.5	0.7	0.8
$nP_{.j}$	2.0	2.0	2.0

Tableau 5

Solution de l'exemple 2

s	$p_1(s)$	$p_2(s)$	$p_3(s)$
1 0 1	0.5	0.4	0.3
1 1 0	0.3	0.2	0.1
0 1 1	0.2	0.1	0.0
1 0 1	0.0	0.1	0.2
1 1 0	0.0	0.1	0.2
0 1 1	0.0	0.1	0.2
1 0 1	0.0	0.1	0.2
1 1 0	0.0	0.1	0.2
0 1 1	0.0	0.1	0.2
1 0 1	0.0	0.1	0.2
1 1 0	0.0	0.1	0.2
0 1 1	0.0	0.1	0.2

Exemple 3: Causey, Cox et Ernst (1985)

Causey et coll. (1985) donnent un exemple d'une stratification tridimensionnelle pour laquelle leur méthode est incapable de donner une solution. Ils examinent une population soumise à une stratification $2 \times 2 \times 2$ de laquelle un échantillon de taille $n = 2$ doit être tiré; les valeurs suivantes sont données pour la taille attendue de l'échantillon dans la ijk ème cellule, n_{ijk} :

$$n_{111} = n_{221} = n_{122} = n_{212} = .5$$
$$n_{121} = n_{211} = n_{112} = n_{222} = 0.$$

Si nous appliquons notre méthode d'une manière semblable à celle employée dans les exemples 1 et 2, nous obtenons la solution donnée au tableau 6. Dans ce cas, la fonction objective n'a pas atteint zéro, de sorte qu'il n'y a pas une correspondance exacte avec les marges dans chaque échantillon.

Tableau 6

Solution de l'exemple 3

s	$i = 1$		$i = 2$		$p(s)$
	1	0	0	1	
1 0 0	0	0	0	0	0.5
0 1 0	0	0	0	0	0.5

4. COMPARAISON DES ERREURS

QUADRATIQUES
MOYENNES

Dans la présente section, l'erreur quadratique moyenne (EQM) du plan proposé avec estimateur \bar{y} sera comparée à celle du plan de Bryant et coll. (1960) pour chacun des deux estimateurs qu'ils proposent, soit \bar{y}_U et \bar{y}_B , où les indices U et B indiquent que le premier estimateur est non biaisé et que le second est biaisé. Désignons les cellules par c (ij dans le cas bidimensionnel), désignons une unité d'une cellule par k (et si nécessaire l), et supprimeons le s dans $n_c(s)$, pour simplifier la notation. La probabilité de sélection de n importe quelle unité k de la cellule c est

(4.1)
$$\pi_{ck} = E[n_c]/N_c = E[n_c]/(NP_c)$$

et la probabilité de sélection conjointe de l'unité k de la cellule c et de l'unité k' de la cellule c' est

(4.2)
$$\pi_{ckc'k'} = \begin{cases} \frac{E[n_c n_{c'}]}{N_c N_{c'}} & \text{si } c \neq c', \\ \frac{E[n_c(n_c-1)]}{N_c(N_c-1)} & \text{si } c = c'. \end{cases}$$

Par conséquent, si $\sigma^2_\alpha = \sigma^2_\beta$, la variance probable de $\bar{y}(s)$ liée au plan selon ce modèle est minimisée par le choix de la fonction de perte définie en (2.9). Autre possibilité: si l'on disposait d'information préalable sur le ratio probable entre la variance entre lignes et la variance entre colonnes, il pourrait être judicieux, pour des besoins d'efficacité, de modifier la fonction de perte décrite en (2.9) en multipliant le premier terme du côté droit de (2.9) par ce ratio estimatif.

Par contre, si l'on pressent que la valeur de Y sera assujettie à une forte interaction entre les facteurs lignes et les facteurs colonnes, il n'est peut-être pas approprié de tenter simplement d'équilibrer l'échantillon selon les marges. Par exemple, si l'un des facteur de stratification est "urbain/rural" et que l'autre est un indicateur économique X , et que l'on sait que Y a un lien positif avec X dans les régions urbaines et un lien négatif dans les régions rurales, il sera sans doute plus efficace de stratifier partiellement selon X *séparément* pour les régions rurales et les régions urbaines, que d'équilibrer complètement selon les deux marges. Voir Bryant et coll. (1960, section 9) pour des commentaires connexes sur l'efficacité dans le cas d'une stratification bidimensionnelle.

2.3 Stratifications de dimensions supérieures

On peut, de façon naturelle, étendre la méthode proposée à trois facteurs de stratification ou plus, en définissant s comme étant la matrice à r dimensions correspondante. La fonction de perte comprendra généralement des termes additionnels; par exemple, pour une stratification à trois dimensions, nous pourrions poser

$$w(s) = \lambda_1 \sum_{i=1}^{R_1} (n_{i..}(s) - nP_{i..})^2 + \lambda_2 \sum_{j=1}^{R_2} (n_{.j.}(s) - nP_{.j.})^2 + \lambda_3 \sum_{k=1}^{R_3} (n_{..k}(s) - nP_{..k})^2$$

selon une notation évidente, où λ_1 , λ_2 et λ_3 sont inclus pour représenter l'importance relative d'un équilibrage selon chacun des trois facteurs, et pourraient être des estimations préalables des variances des moyennes de Y entre les catégories des trois facteurs de stratification, comme en (2.10).

2.4 Échantillonnage à plusieurs degrés

Une importante application pratique de la stratification multidimensionnelle consiste à sélectionner les unités primaires d'échantillonnage (upé) d'un échantillon à plusieurs degrés, car il est fréquent qu'on dispose d'information sur plusieurs variables de stratification.

3. EXEMPLES

Exemple 1: Bryant, Hartley et Jessen (1960)

Dans la méthode exposée à la section 2.1, les probabilités de sélection de chaque unité de la population sont $E(n_{ij}(s)/N_{ij}) = n/N$. Si l'on désire choisir des upé avec probabilité égale, on peut appliquer immédiatement cette méthode en considérant les upé comme les unités et en remplaçant les valeurs observées de Y par des estimateurs sans biais des totaux pour les upé. Supposons toutefois que nous souhaitions choisir des upé avec probabilités inégales, par exemple n_{ijk} pour l'upé k dans la cellule ij , où généralement n_{ijk} sera égal à $M_{ijk}/\sum_{jk} M_{ijk}$, M_{ijk} étant une quelconque mesure de la taille de l'upé k dans la cellule ij . On peut alors modifier simplement le processus, en posant P_{ij} égal à la somme des z_{ijk} pour les upé k de la cellule ij . Alors, si $n_{ij}(s) > 0$, un échantillon de $n_{ij}(s)$ upé est prélevé dans la cellule ij selon une méthode dans laquelle la probabilité est proportionnelle à z_{ijk} .

Nous allons d'abord appliquer la méthode à l'exemple hypothétique de Bryant et coll. (1960) donné au tableau 1. Nous commençons par ramener le problème à la forme décrite en (2.4), (2.5) et (2.6), avec les r_{ij} donnés au tableau 2. La fonction de poids présentée en (2.9) devient, dans ce problème réduit de programmation linéaire,

$$w(s) = \sum_5 (\bar{n}_{i.}(s) - r_{i.})^2 + \sum_3 (\bar{n}_{.j}(s) - r_{.j})^2.$$

A l'aide d'un logiciel courant de programmation linéaire de la bibliothèque NAG FORTRAN, nous obtenons la solution donnée au tableau 3. Les valeurs I_{ij} ont été ajoutées à la solution de telle façon que $n_{ij} = I_{ij} + \bar{n}_{ij}(s)$. Il se révèle, pour cette solution, que chaque s pour lequel $p(s) > 0$, a des marges $n_{i.}(s)$ et $n_{.j}(s)$ qui correspondent exactement aux marges désirées, c'est-à-dire que la solution rend (2.4) égal à zéro.

Tableau 3
Solution de l'exemple 1

s			$d(s)$	s	$d(s)$
0.1			1	1	1
			0	1	0
			1	1	1
			0	1	1
			0	1	0
			1	0	1
			1	0	1
			1	1	0
0.2			1	1	0
			0	2	0
			0	0	2
			1	1	0
			1	0	1
			0	1	0
			1	0	1
			1	1	0
0.1			1	1	0
			0	0	1
			1	1	1
			1	1	1
			0	0	1
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0
			1	1	0

$$(2.4) \quad \underset{p \in P}{\text{minimiser}} \sum_{s \in S_n} w(s) p(s),$$

sous les contraintes

$$(2.5) \quad \sum_{s \in S_n} n_{ij}(s) p(s) = r_{ij},$$

$$(2.6) \quad \sum_{s \in S_n} p(s) = 1, \quad 0 \leq p(s) \leq 1 \text{ pour tout } s \in S_n,$$

où S_n est l'ensemble des matrices $R \times C$ où tous les éléments sont 0 ou 1 et où la somme des éléments est $\bar{n} = n - \sum_{ij} I_{ij}$. On peut constater, évidemment, que si tous les I_{ij} sont égaux à zéro, le problème est exactement le même qu'avant. Le nombre d'éléments dans S_n , qui détermine l'ampleur du travail de calcul de la programmation linéaire, est maintenant $\binom{n}{RC}$. Ce nombre peut encore être très élevé, toutefois, et peut être réduit encore davantage par un choix approprié de la fonction de perte $w(s)$, comme nous le verrons dans la section suivante.

Dans l'exemple du tableau 1, on ramènerait la situation à celle représentée au tableau 2, tout en ne permettant que des tailles de cellule égales à 0 ou 1, et en ajoutant ensuite 1 aux cellules (1,1), (3,3), (4,2) et (5,1) dans la solution finale. Ainsi, $n = 10$, mais $\bar{n} = 6$.

Tableau 2

Tableau des valeurs r_{ij} , d'après le tableau 1, avec $\bar{n} = 6$

Région	Type de localité			
	Urbaine	Rurale	Métropolitaine	Total
1	0.0	0.5	0.5	1.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	0.2	1.0
4	0.6	0.8	0.6	2.0
5	0.0	0.8	0.2	1.0
Total	1.0	3.0	2.0	6.0

2.2 Choix de la fonction de perte $w(s)$

La méthode proposée est particulièrement souple du fait qu'elle laisse à l'utilisateur le choix de la fonction $w(s)$ incluse dans l'énoncé de la fonction objective en (2.2). La façon de procéder habituelle, dans le cas d'une stratification bidimensionnelle (p. ex. Jessen 1970; Causey et coll. 1985), consiste à exiger que l'échantillon prélevé respecte les contraintes marginales:

$$(2.7) \quad |n_{i\cdot}(s) - nP_{i\cdot}| > 1 \quad i = 1, \dots, R,$$

$$(2.8) \quad |n_{\cdot j}(s) - nP_{\cdot j}| > 1 \quad j = 1, \dots, C,$$

où

$$n_{i\cdot}(s) = \sum_{j=1}^C n_{ij}(s), \quad n_{\cdot j}(s) = \sum_{i=1}^R n_{ij}(s)$$

$$P_{i\cdot} = \sum_{j=1}^C P_{ij}, \quad P_{\cdot j} = \sum_{i=1}^R P_{ij}.$$

Si, toutefois, nous utilisons dans notre méthode une fonction de perte comme

$$(2.9) \quad w(s) = \sum_{i=1}^R (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^C (n_{\cdot j}(s) - nP_{\cdot j})^2,$$

il existera toujours une solution optimale dans un ensemble S_n suffisamment grand. Il peut être avantageux, du point de vue des besoins de calcul, de restreindre au départ l'ensemble S_n aux seuls échantillons respectant les contraintes (2.7) et (2.8), ou même à un sous-ensemble de ces derniers, puis d'élargir l'ensemble, si nécessaire, en remplaçant par exemple 1 par 2 dans (2.7) et (2.8), jusqu'à ce qu'une solution soit trouvée.

Examinons maintenant un aspect plus fondamental et voyons en quoi des contraintes comme (2.7) et (2.8) sont pertinentes. D'un point de vue non statistique, l'équilibrage d'un échantillon par rapport à des facteurs ayant une distribution connue pour l'ensemble de la population peut rassurer les utilisateurs quant à la "représentativité" de l'échantillon. D'un point de vue statistique, compte tenu de notre contrainte d'une estimation sans biais (2.1), il est naturel de se demander comment choisir la fonction de perte de façon à améliorer l'efficacité. On peut examiner cette question en posant $w(s)$ comme étant l'erreur quadratique moyenne $E_m(\bar{y}(s) - \bar{Y})^2$ en vertu d'un modèle m . Dans ce cas, la solution du problème d'optimisation (2.2) minimise l'espérance, selon le plan, de l'erreur quadratique moyenne liée au modèle ou, de façon équivalente puisque nous exigeons une estimation non biaisée selon le plan, l'espérance, selon le modèle, de la variance liée au plan.

Considérons, par exemple, le modèle d'analyse de variance des effets principaux

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

où y_{ijk} est la k ème valeur de Y dans la cellule ij , μ est une moyenne fixe et α_i , β_j et ϵ_{ijk} sont des effets aléatoires indépendants de moyenne zéro et de variances σ_α^2 , σ_β^2 et σ_ϵ^2 respectivement. On a alors, sans tenir compte des termes de correction pour population finie,

$$E_m(\bar{y}(s) - \bar{Y})^2 = \sigma_\alpha^2 \sum_{i=1}^R (n_{i\cdot}(s)/n - P_{i\cdot})^2$$

$$+ \sigma_\beta^2 \sum_{j=1}^C (n_{\cdot j}(s)/n - P_{\cdot j})^2 + \sigma_\epsilon^2/n. \quad (2.10)$$

Toutes les méthodes ci-dessus peuvent être appliquées manuellement avec un degré d'effort variable, mais aucune ne tire parti du potentiel de l'informatique moderne. Dans cet article, nous allons montrer comment les techniques informatiques de la programmation linéaire peuvent être appliquées au problème de la stratification multidimensionnelle, comme l'ont fait auparavant Rao et Nigam (1990, 1992). La méthode que nous proposons peut être considérée comme un complément de la méthode de programmation linéaire proposée par Caussey et coll. (1985). Le choix entre les deux dépendra de la nature du problème de stratification et du logiciel disponible. Notre méthode a comme inconvénient possible d'être beaucoup plus exigeante en calculs, puisque le nombre d'inconnues dans notre problème de programmation linéaire peut être aussi élevé que $\binom{n}{k}$, où k est le nombre de cellules du tableau et n est la taille de l'échantillon, tandis qu'il n'y a que k inconnues dans la méthode de Caussey et coll. (1985). Nous ferons toutefois quelques suggestions destinées à réduire les calculs exigés par notre méthode. Il s'agit d'une méthode qui est susceptible d'offrir plusieurs avantages. Premièrement, la correspondance est directe entre le problème de stratification et le problème de programmation linéaire, de sorte que la programmation informatique est immédiate, tandis que la méthode de Caussey et coll. est moins directe, ce qui oblige à simuler le comportement de fonctions non linéaires à l'aide de fonctions linéaires (p. 904) et à imbriquer des problèmes de programmation linéaire répétés dans un algorithme comportant une récursivité additionnelle. Deuxièmement, notre méthode donne toujours une solution, tandis que celle de Caussey et coll. n'en produit pas toujours une, par exemple dans des cas de stratification tridimensionnelle. Troisièmement, la fonction objective de notre problème de programmation linéaire peut être naturellement modifiée en fonction de différents objectifs du problème de stratification, par exemple pour une stratification tridimensionnelle dans laquelle il est plus important d'"équilibrer" l'échantillon à l'égard des deux premières variables de stratification qu'à l'égard de la troisième. Quatrièmement, notre méthode peut être naturellement modifiée pour forcer les probabilités d'inclusion conjointe de cellules à être positives, de façon à permettre une estimation non biaisée de la variance.

2. MÉTHODE PROPOSÉE

2.1 Concepts de base

Examinons, pour commencer, la forme la plus simple de stratification bidimensionnelle. Soit une population de N unités répartie entre les RC cellules d'un tableau bidimensionnel résultant de la classification croisée d'un facteur colonne comprenant R catégories et d'un facteur ligne comprenant C catégories. Soit N_{ij} le nombre d'unités de la cellule ij , c'est-à-dire de l'ensemble d'unités appartenant à la fois à la ligne i et à la colonne j , et soit $P_{ij} = N_{ij}/N$ la proportion correspondante. Supposons que le paramètre d'intérêt soit la moyenne de la population, \bar{Y} , d'une variable Y .

Nous limitons aussi notre attention à une stratification proportionnelle telle que

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij}(s) = n.$$

Nous limitons notre examen aux plans comportant une taille d'échantillon fixe $n > 0$, c'est-à-dire que nous restreignons S à l'ensemble S_n de toutes les matrices telles que

Examinons la méthode d'échantillonnage à deux degrés suivantes. Premièrement, des tailles d'échantillon n_{ij} sont déterminées pour chaque cellule d'après une procédure aléatoire donnée. Si s désigne la matrice $R \times C$ des $(n_{ij}, i = 1, \dots, R, j = 1, \dots, C)$, cette procédure attribue une probabilité $p(s)$ à chaque s d'un ensemble de S matrices possibles. Pour bien montrer que les n_{ij} dépendent de s , nous écrivons $n_{ij}(s)$. Deuxièmement, un échantillon aléatoire simple de $n_{ij}(s)$ unités est prélevé dans la cellule ij et les valeurs de Y propres aux unités de l'échantillon sont enregistrées.

Il résulte de (2.1) que la moyenne non pondérée simple de l'échantillon, $\bar{y}(s)$, est un estimateur sans biais de \bar{Y} . Nous proposons de choisir un (ou le) plan d'échantillonnage $p(s)$ qui minimise le degré de "non-désirabilité" attendu de l'échantillon s , en résolvant le problème:

$$(2.2) \quad \text{minimiser}_{s \in S_n} \sum_{p \in P} w(s)p(s),$$

sous la contrainte (2.1), où $w(s)$ est une fonction de perte (à préciser) qui s'applique à l'échantillon s , et P est la classe des plans d'échantillonnage possibles sur S_n qui remplissent la condition

$$(2.3) \quad 0 \leq p(s) \leq 1 \quad \text{pour tout } s \in S_n.$$

Notons qu'il découle de (2.1) que $\sum_{s \in S_n} p(s) = 1$. L'observation clé de Rao et Nigam (1990, 1992) est que la fonction objective énoncée en (2.2), ainsi que les contraintes d'égalité et d'inégalité énoncées en (2.1) et (2.3), sont toutes des fonctions linéaires en $p(s)$, et donc qu'on peut résoudre ce problème directement par programmation linéaire, les inconnues étant $p(s)$, $s \in S_n$. Le principal obstacle à l'application de cette méthode vient du fait que le nombre d'éléments de S_n est souvent très élevé et qu'il est difficile, même avec les ressources informatiques modernes, de résoudre par programmation linéaire un problème qui comporte un grand nombre d'inconnues.

Il vaut mieux, par conséquent, restreindre l'examen à un sous-ensemble de S_n . Une restriction naturelle consiste à ne tenir compte que des matrices s pour lesquelles $n_{ij}(s)$ est égal soit à $I_{ij} = [n P_{ij}]$, le plus grand entier inférieur à $n P_{ij}$, soit à $I_{ij} + 1$. Si l'on pose $n_{ij}(s) = n_{ij}(s) - I_{ij}$ et $r_{ij} = n P_{ij} - I_{ij}$, le problème devient

Stratification multidimensionnelle par programmation linéaire

R.R. SITTER et C.J. SKINNER¹

RÉSUMÉ

Rao et Nigam (1990, 1992) ont montré comment une classe de plans à échantillonnage contrôle peut être produite par programmation linéaire. Dans cet article, leur méthode est appliquée à la stratification multidimensionnelle. L'examen des plans de sondage produits dans des applications particulières, ainsi que l'évaluation des erreurs quadratiques moyennes, permettent d'établir une comparaison avec les méthodes existantes. La méthode proposée est d'utilisation relativement simple et semble avoir une performance raisonnable sur le plan des erreurs quadratiques moyennes. Les calculs nécessaires peuvent toutefois augmenter rapidement, à mesure que s'accroît le nombre de cellules dans la classification multidimensionnelle. L'estimation de la variance est également examinée.

MOTS CLÉS: Choix contrôle; programmation linéaire; échantillonnage à plusieurs degrés; échantillonnage stratifié.

1. INTRODUCTION

Souvent, le concepteur d'une enquête dispose de plusieurs variables pour faire une stratification et choisit naturellement de définir les strates comme étant les cellules formées par la classification multidimensionnelle de ces variables. Un problème que pose cette méthode, notamment lorsqu'on prélève les unités primaires d'échantillonnage (upé) pour des enquêtes sur les ménages, vient du fait que la taille désirée de l'échantillon peut être inférieure au nombre total de cellules, auquel cas les méthodes de stratification classiques peuvent ne pas s'appliquer.

Une illustration, basée sur un exemple hypothétique de Bryant et coll. (1960), est donnée au tableau 1. Des localités (upé) sont classées selon deux variables de stratification: le type de localité (trois catégories) et la région (cinq catégories). La taille désirée de l'échantillon, $n = 10$, est inférieure au nombre total de cellules, 15. Cet exemple fait apparaître un problème connexe. Les valeurs du tableau 1 sont les nombres attendus selon une stratification proportionnelle, c'est-à-dire les proportions de la population multipliées par la taille de l'échantillon. Même si la taille de l'échantillon était doublée, afin qu'elle dépasse le nombre de cellules, les nombres attendus ne seraient toujours pas des entiers. Le fait d'arrondir ces valeurs à des nombres entiers aurait sans doute un effet pratiquement négligeable pour des nombres élevés, mais le choix de la méthode d'arrondissement est plus préoccupant lorsque ces nombres sont très faibles.

Une solution au problème que pose le trop grand nombre de cellules consiste à laisser de côté une ou plusieurs variables de stratification, ou encore à fusionner certaines catégories. Ont également été proposées diverses autres méthodes en vertu desquelles on tente de maintenir un certain "contrôle" pour toutes les catégories de l'ensemble des variables de stratification, en permettant des formes différentes de sélection aléatoire des cellules.

Tableau 1

Région	Type de localité			Total
	Urbaine	Rurale	Métropolitaine	
1	1.0	0.5	0.5	2.0
2	0.2	0.3	0.5	1.0
3	0.2	0.6	1.2	2.0
4	0.6	1.8	0.6	3.0
5	1.0	0.8	0.2	2.0
Total	3.0	4.0	3.0	10.0

Nombres probables d'unités de l'échantillon dans les cellules en vertu d'une stratification proportionnelle avec $n = 10$

Goodman et Kish (1950) ont proposé une méthode qu'ils ont baptisée "choix contrôle". Jessen (1970) se dit d'avis que "cette méthode est un peu complexe et son utilisation en échantillonnage appliqué apparaît limitée" (p. 778). Waterton (1983) illustre cette complexité, tandis que Bryant et coll. (1960) proposent une méthode beaucoup plus simple pour une stratification bidimensionnelle. Cette méthode se caractérise par une indépendance des nombres attendus d'unités échantillonnées entre les lignes et les colonnes du tableau bidimensionnel. Si les lignes et les colonnes sont également indépendantes dans la population, aucun problème ne se pose, mais si l'on note un degré important de dépendance (comme c'est souvent le cas), il faut habituellement revoir la pondération, ce qui peut rendre cette solution intéressante en pratique et accroître la variance, comme il est démontré à la section 5. Jessen (1970) signale qu'une autre limite de la méthode de Bryant et coll. (1960) vient de l'impossibilité de contraindre des cellules déterminées à avoir une taille nulle. Il propose deux méthodes, applicables à des stratifications bidimensionnelles et tridimensionnelles, mais ces méthodes demeurent d'application assez difficile et, comme le notent Causey et coll. (1985), elles ne produisent pas toujours une solution.

¹ R.R. Sitter, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6; C.J. Skinner, Department of Social Statistics, University of Southampton S09 5NH, U.K.

REMERCIEMENTS

Les auteurs remercient le professeur J.N.K. Rao de leur avoir suggéré l'étude de ce problème, et sont reconnaissants envers l'arbitre pour les suggestions constructives formulées. Cette recherche a bénéficié de subventions du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

- KULLDORFF, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Revue de l'Institut International de Statistique*, 31, 24-57.
- MURTHY, N.N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.
- PRASAD, N.G.N., et SRIVENKATARAMANA, T. (1980). Double sampling with PPS selection. *Vignana Bharathi*, 6, 52-58.
- RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.
- RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.
- SUKHATME, P.V., et SUKHAATME, B.V. (1970). *Sampling Theory of Surveys With Applications*. Ames, Iowa: Iowa State University Press.
- CHOTAL, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Série C, 36, 173-180.
- GHANGURDE, P., et RAO, J.N.K. (1969). Some results on sampling over two occasions. *Sankhyā*, Série A, 31, 463-472.

3. COMPARAISONS NUMÉRIQUES DE L'EFFICACITÉ

Les estimateurs composites Y_C^2 défini en (1.12), Y_{CM}^2 défini en (1.16) et Y_C^2 défini en (2.1) sont maintenant comparés à leurs valeurs optimales Q et λ respectives. L'efficacité du plan proposé en 2.2 par rapport à la méthode de Chotali (1974) est examinée au moyen d'une comparaison des deux efficacités relatives suivantes:

$$REI = \frac{V_{min}(Y_C^2)}{V_{min}(Y_C^2)} = \frac{(1 - n/N) + \sqrt{2(1 - \delta)}}{(1 - n/N) + \sqrt{h}}$$

et

$$RE2 = \frac{V_{min}(Y_{CM}^2)}{V_{min}(Y_C^2)} = \frac{(1 - n/N) + \sqrt{1 - \delta^2}}{(1 - n/N) + \sqrt{h}}$$

obtenues respectivement à l'aide de (1.15) et (2.4), et de (1.19) et (2.4). Il s'ensuit que le plan proposé est supérieur au plan de Chotali utilisant un estimateur de Kulldorff (qui dépend de la constante inconnue β) dans le cas des populations pour lesquelles $h > (1 - \delta^2)$. Afin de permettre des comparaisons numériques utiles, nous utilisons deux ensembles de données déjà examinés dans la littérature.

Ensemble de données A: Cet ensemble de données a trait à la superficie de culture du blé en 1964 (Y_2) et en 1963 (Y_1), ainsi qu'à la superficie cultivée en 1961 (x), pour 34 villages de l'Inde (voir Murthy 1967). Les valeurs des paramètres pour cet ensemble de données sont $\delta = 0.6404$ et $h = 0.1868$.

D'après ces valeurs de δ et h , les deux efficacités relatives RE1 et RE2 (exprimées en pourcentage) ont été calculées pour certaines valeurs de n/N , et les résultats sont présentés aux tableaux 1 et 2.

Tableau 1

REI en % pour les ensembles de données A et B		
n/N	Ensemble A	Ensemble B
0.05	130.09	124.30
0.10	131.22	125.21
0.15	132.43	126.19
0.20	133.75	127.25
0.25	135.18	128.41
0.30	136.73	129.66

Un examen du tableau 1 permet de conclure que le plan proposé est supérieur à celui de Chotali (1974). Le gain d'efficacité passe de 30% à 37% pour l'ensemble de données A et de 24% à 30% pour l'ensemble de données B avec l'augmentation de 0.05 à 0.30 de la fraction de sondage. Notons que le gain d'efficacité est plus élevé pour l'ensemble de données A que pour l'ensemble de données B en raison de la différence entre les valeurs des paramètres h (0.1868 contre 0.3811) et δ (0.6404 contre 0.7635). Rappelons que h évalue l'efficacité de p_i en tant que mesure de la taille pour l'unité i par rapport à celle de Y_{1i} en tant que mesure de la taille dans l'estimation du total Y_2 pour l'occasion courante, tandis que δ est la corrélation entre Y_{1i}/p_i et Y_{2i}/p_i telle que définie en (1.7). Quand h est relativement petit, on obtient avec la méthode proposée des gains d'efficacité supérieurs à ceux obtenus quand h est élevé. Dans les deux cas, toutefois, les gains d'efficacité que procure la méthode proposée sont dignes d'intérêt.

Les gains d'efficacité qu'engendre la méthode proposée par rapport au plan de Chotali utilisant un estimateur de Kulldorff (résultats présentés au tableau 2) sont minimes, variant de 4.5% à 5.3% pour l'ensemble de données A et de 1.8% à 2.2% pour l'ensemble de données B. Toutefois, il est nécessaire de connaître la valeur de β pour pouvoir utiliser l'estimateur de Kulldorff, ce qui n'est pas le cas en pratique. La stratégie proposée, donc, donne de bons résultats tant du point de vue de son application en situation réelle que de celui du gain d'efficacité.

Il existe des situations dans lesquelles l'information auxiliaire nécessaire au calcul des probabilités de sélection initiales n'est pas disponible. Un échantillonnage aléatoire simple peut alors être utilisé en remplacement de la méthode RHC pour la sélection de l'échantillon de la première occasion; on peut ensuite appliquer la méthode RHC au prélèvement de s_i , en utilisant l'information sur la variable à l'étude fournie par l'échantillonnage aléatoire simple effectué à la première occasion. La théorie relative à une telle méthode s'obtient directement à titre de cas spécial de celui présenté, en prenant $p_i = 1/N$, $i = 1, \dots, N$. D'importants gains d'efficacité devraient être obtenus dans ce cas.

Tableau 2

RE2 en % pour les ensembles de données A et B		
n/N	Ensemble A	Ensemble B
0.05	104.49	101.82
0.10	104.64	101.88
0.15	104.80	101.94
0.20	104.97	102.01
0.25	105.15	102.08
0.30	105.34	102.16

Examinons maintenant un estimateur du total pour la deuxième occasion, Y_2 , qui exploite la méthode proposée.

Posons

$$y_{2i}^* = \frac{y_{2i} P_i}{d_i}.$$

Un estimateur composite de Y_2 est donné par

$$Y_2^* = \bar{Q}^{**} Y_{2n}^C + (1 - \bar{Q}^{**}) Y_{2m}^*, \quad (2.1)$$

où Y_{2n}^C est défini comme en (1.13), $0 \leq \bar{Q}^{**} \leq 1$ et

$$Y_{2m}^* = \sum_{i=1}^l y_{2i}^* \bar{P}_i^*.$$

Ici, \bar{P}_i^* désigne le total des valeurs P_i^* associées aux unités appartenant au groupe aléatoire d'où provient la i ème unité de s_1 . Supposons que E_1 et E_2 désignent l'espérance et que V_1 et V_2 désignent la variance sur tous les s et pour un s donné, respectivement. On peut alors constater que Y_{2m}^* , et donc Y_2^* pour Y_2 , sont sans biais, car l'espérance de Y_{2m}^* est

$$E(Y_{2m}^*) = E_1 E_2(Y_{2m}^*) = E_1 \left(\sum_{i=1}^l \frac{y_{2i} P_i}{d_i} \right) = Y_2. \quad (2.2)$$

Pour obtenir la variance de Y_{2m}^* , considérons

$$V_2(Y_{2m}^*) = \frac{n-m}{n-m} \sum_{i=1}^l \left(\frac{y_{2i}^*}{d_i^*} \right)^2 \bar{P}_i^* - \sum_{i=1}^l \frac{y_{2i}^*}{d_i^*} \sum_{j=1}^l \frac{y_{2j}^*}{d_j^*} \bar{P}_i^* \bar{P}_j^*.$$

$$= \frac{n-m}{n-m} \sum_{i=1}^l \left[\sum_{j=1}^l \frac{y_{2i}^* y_{2j}^*}{d_i^* d_j^*} \bar{P}_i^* \bar{P}_j^* \right] - \sum_{i=1}^l \frac{y_{2i}^*}{d_i^*} \sum_{j=1}^l \frac{y_{2j}^*}{d_j^*} \bar{P}_i^* \bar{P}_j^*.$$

$$- \sum_{i=1}^l \left(\sum_{j=1}^l \frac{y_{2i}^* y_{2j}^*}{d_i^* d_j^*} \right)^2 \bar{P}_i^* \bar{P}_j^*,$$

ce qui donne, après une importante simplification algébrique

$$E_1 V_2(Y_{2m}^*) = \frac{N(n-m)}{N(n-m)} \sigma_2^2,$$

où

$$\sigma_2^2 = \sum_{i=1}^l \left(\frac{y_{2i}^*}{d_i^*} \right)^2 Y_1 - \frac{Y_1^2}{d_i^*}.$$

Notons que la quantité h reflète l'efficacité de l'estimateur utilisant les P_i comme probabilités de sélection initiales par rapport à celle de l'estimateur fondé sur les probabilités de sélection initiales y_{1i}/X_1 . Une valeur "peu élevée" de h entraîne un accroissement de l'efficacité de la méthode proposée par rapport à celle de Chotal.

on a

$$V_1 E_2(Y_{2m}^*) = \frac{N-n}{N-n} \sigma_2^2, \quad (2.3)$$

Puisque

$$V_1 E_2(Y_{2m}^*) = \frac{N-n}{N-n} \sigma_2^2,$$

Parce que Y_{2n}^C et Y_{2m}^* sont indépendants, la variance de Y_2^* est donnée par

$$V(Y_2^*) = \bar{Q}^{**2} V(Y_{2n}^C) + (1 - \bar{Q}^{**})^2 V(Y_{2m}^*),$$

où

$$V(Y_{2n}^C) = \frac{N-n}{N-n} \sigma_2^2,$$

et $V(Y_{2m}^*)$ est donnée par (2.3).

On obtient la variance minimum de $V(Y_2^*)$ en utilisant les valeurs optimales de \bar{Q}^{**} et λ , données respectivement

par

$$\bar{Q}^{**} = \frac{(1 - n/N)h + \frac{\lambda}{(1-\lambda)}}{(1 - n/N)h + \frac{\lambda}{(1-\lambda)}},$$

et

$$\lambda = \frac{1 + \sqrt{h}}{\sqrt{h}}.$$

Donc, la variance minimum de $V(Y_2^*)$ est donnée par

$$V^{min}(Y_2^*) = \frac{N\sigma_2^2}{n(N-1)} [1 - n/N + \sqrt{h}]. \quad (2.4)$$

En vertu de ce plan, mais sans l'hypothèse (1.5), Chotai a aussi examiné un estimateur de Y_2 (semblable à l'estimateur de Kulldorff pour l'échantillonnage aléatoire simple; voir Kulldorff 1963) donné par

$$Y_{CM}^2 = \bar{Q}_{CM} Y_{2u}^2 + (1 - \bar{Q}_{CM}) Y_{2m}^2, \quad (1.16)$$

où Y_{2u}^2 est défini comme en (1.13), \bar{Q}_{CM} ($0 \leq \bar{Q}_{CM} \leq 1$) est un coefficient de pondération donné qui doit être déterminé et

$$Y_{2m}^2 = \sum_{i=1}^m \frac{Y_{2i}^2 - \beta Y_{1i}^2 P_i^2}{Y_{1i}^2 P_i^2} + \beta \sum_{i=1}^m \frac{P_i^2}{Y_{1i}^2 P_i^2}, \quad (1.17)$$

avec

$$\beta = \delta = \frac{\sum_{i=1}^m P_i^2 (Y_{2i}^2 / P_i^2 - Y_2^2)}{\sum_{i=1}^m P_i^2 (Y_{1i}^2 / P_i^2 - Y_1^2)^2} = \delta \frac{V_1^2}{V_2^2} \quad (1.18)$$

et δ tel que défini en (1.7). La variance minimum de Y_{CM}^2 , selon les valeurs optimales de \bar{Q}_{CM} et λ , est donnée par

$$V_{min}^2(Y_{CM}^2) = \frac{2n(N-1)}{N} (1 + \sqrt{1 - \delta^2} - n/N) V_2. \quad (1.19)$$

Pour utiliser effectivement Y_{CM}^2 , il est évidemment nécessaire de trouver d'abord la valeur de β , ce qui, en général, n'est pas possible en pratique. On peut utiliser une estimation de β fondée sur l'échantillon disponible, mais cela introduira un biais dans l'estimateur Y_{CM}^2 .

2. AUTRES PLANS POSSIBLES POUR UN ÉCHANTILLONNAGE AVEC PPT EN DEUX OCCASIONS

Nous présentons maintenant une autre méthode possible d'échantillonnage et d'estimation, qui n'exige pas de connaître la valeur β telle que définie en (1.18). Selon ce plan, on se sert d'une information recueillie à la deuxième occasion pour sélectionner l'échantillon à la deuxième occasion. La méthode se fonde sur une technique que Prasad et Srivenkataramana (1980) ont élaborée dans le contexte d'un échantillonnage double dans lequel un sous-échantillon de deuxième phase est prélevé selon l'information obtenue d'un échantillon initial. À des fins de simplicité, nous examinons d'abord son application au plan de Raj (1965) décrit plus haut.

2.1 Modification du plan de Des Raj

À la première occasion, un échantillon s de taille n est sélectionné avec probabilités p_i proportionnelles aux valeurs x_i et avec remplacement. À la deuxième occasion, plutôt que de prélever un sous-échantillon aléatoire simple sans remplacement, on prélève un sous-échantillon s_1 de m unités dans s , selon un plan avec PPT et avec remplacement, en prenant comme mesure de la taille $z_i = y_{1i}/x_i$, où y_{1i} est la valeur observée de la caractéristique y pour l'unité i à la première occasion. Un échantillon s_2 de taille $u = n - m$ est tiré indépendamment de s , comme dans Raj (1965). Un estimateur composite de Y_2 est donné par

$$Y_2 = \bar{Q} Y_{2u} + (1 - \bar{Q}) Y_{2m},$$

où Y_{2u} est tel que défini en (1.3) et

$$Y_{2m} = \frac{1}{m} \sum_{i=1}^m \frac{(Y_{2i}^2 / P_i^2)}{(Y_{1i}^2 / P_i^2)} \sum_{i=1}^m (Y_{1i}^2 / P_i^2),$$

\bar{Q} étant un coefficient de pondération, $0 \leq \bar{Q} \leq 1$. La variance minimum de Y_2 , obtenue en minimisant la variance de Y_2 par rapport à \bar{Q} , est donnée par

$$V_{min}^2(Y_2) = V_1 C_1 (n + C_1 m)^{-1},$$

où $C_1 = \sum_{i=1}^n (Y_{2i}^2 / P_i^2 - Y_2^2) P_i^2 V_1^{-1}$, avec $P_{1i} = y_{1i} / Y_1$ et V_1 tel que défini en (1.5).

2.2 Modification du plan de Chotai

Supposons, comme dans Chotai (1974), que N , n et m ($m < n$) soient tous des entiers positifs tels que N/n , N/m et n/m sont aussi des entiers. Alors:

1. À la première occasion, prélevons un échantillon s de taille n selon une méthode identique à celle du plan G-R. Pour cet ensemble d'unités, des observations y_{1i} , $i = 1, \dots, n$, sont faites au sujet d'une caractéristique y .

2. À la deuxième occasion, a) répartissons les n unités de s au hasard en m groupes de taille n/m et tirons indépendamment une unité avec PPT, $p_i^* = (y_{1i} P_i^2) / P_i$, de chacun des m groupes afin de constituer un sous-échantillon s_1 , où P_i^* est tel que défini dans le plan G-R; b) sélectionnons s_2 , un échantillon entièrement nouveau de $u = n - m$ unités dans la population entière, et observons les valeurs y de la deuxième occasion, y_{2i} , pour ces u unités de la même manière que dans le plan G-R.

Notons que la différence entre la méthode proposée et celle de Chotai (1974) réside dans la sélection de s_1 dans le premier cas, on se sert de l'information recueillie à la première occasion pour prélever s_1 à la deuxième occasion.

on trouve que la variance minimum de Y_2 est

$$V_{min}^2(Y_2) = V[1 + \sqrt{2(1 - \delta)/(2n)}], \quad (1.6)$$

où δ est donné par

$$V\delta = \sum_{i=1}^N (y_{1i}/p_i - Y_1)(y_{2i}/p_i - Y_2)p_i. \quad (1.7)$$

1.2 Le plan de Changurde-Rao (G-R)

En utilisant un plan d'échantillonnage avec PPT sans

remplacement, Changurde et Rao (1969) ont étendu la

“méthode de Rao-Hartley-Cochran (RHC)”, aussi appelée

“méthode des groupes aléatoires” (voir Rao, Hartley et

Cochran 1962) à l'échantillonnage en deux occasions.

Dans la méthode RHC, la population de N unités est

divisée au hasard en n groupes de tailles N_1, N_2, \dots, N_n

telles que $\sum_{h=1}^n N_h = N$, et un échantillon d'une unité

est tiré indépendamment de chacun des n groupes avec

probabilités proportionnelles aux probabilités de sélection

initiales, p_i . Dans la méthode G-R, la population est

d'abord divisée au hasard en n groupes de taille N/n

(qu'on suppose être un entier). À la première occasion, une

unité est tirée de chaque groupe aléatoire (comme ci-

dessus), ce qui produit un échantillon s de n unités. À la

deuxième occasion, un échantillon aléatoire simple s_1

de $m = \lambda n (0 < \lambda < 1)$ unités apparées est tiré de s

sans remplacement, et un échantillon indépendant s_2 de

$u = n - m$ unités est tiré de l'ensemble de la population

de N unités par la même méthode que celle utilisée pour

produire s . On obtient alors un estimateur composite de

Y_2 sous la forme suivante

$$Y_2' = \bar{Q}' Y_{2n}' + (1 - \bar{Q}') Y_{2m}', \quad (1.8)$$

où $0 \leq \bar{Q}' \leq 1$,

$$Y_{2n}' = \sum_{i \in s_2} y_{2i} p_i^*, \quad (1.9)$$

et

$$Y_{2m}' = \sum_{i \in s_1} y_{1i} p_i^* + nm^{-1} \sum_{i \in s_1} (y_{2i} - y_{1i}) p_i^*, \quad (1.10)$$

P_i et P_i^* désignant les totaux des valeurs p_i pour les groupes contenant la i ème unité ($i = 1, 2, \dots, N$) dans la sélection de s et s_2 respectivement. En vertu de l'hypothèse (1.5), la variance de Y_2' (avec valeurs optimales de \bar{Q}' et λ) est donnée par

$$NV_{min}^2(Y_2') = \frac{2n(N - 1)}{NV}$$

$$\times [1 - n/N + \sqrt{2(1 - \delta)(1 + \gamma)n/N}], \quad (1.11)$$

Chotai (1974), en posant comme hypothèse additionnelle que n/m est un entier, a modifié le plan d'échantillonnage G-R pour la deuxième occasion. Un échantillon s est prélevé comme dans le plan G-R à la première occasion. À la deuxième occasion, les n unités de l'échantillon s sont réparties au hasard en $m (= \lambda n)$ groupes de taille n/m . Une unité est tirée indépendamment de chacun des m groupes avec probabilités proportionnelles aux P_i , tels que définis dans le plan G-R. Cette sélection donne l'échantillon s_1 . La sélection de s_2 se fait de la même façon que dans le plan G-R. On dispose alors d'un estimateur composite de Y_2 sous la forme

$$Y_2' = \bar{Q} Y_{2n}' + (1 - \bar{Q}) Y_{2m}', \quad (1.12)$$

où $0 \leq \bar{Q} \leq 1$,

$$Y_{2n}' = \sum_{i \in s_2} y_{2i} P_i^*, \quad (1.13)$$

et

$$Y_{2m}' = \sum_{i \in s_1} (y_{2i} - y_{1i}) P_i^* + \sum_{i \in s_1} y_{1i} P_i^*. \quad (1.14)$$

Ici, P_i et P_i^* sont définis comme dans le plan G-R, et P_i^+ désigne le total des valeurs P_i pour les groupes aléatoires de s contenant la i ème unité ($i = 1, 2, \dots, N$) dans la sélection de s_1 . La variance minimum de Y_2' en vertu de l'hypothèse (1.5), obtenue à l'aide des valeurs optimales de \bar{Q} et λ , est donnée par

$$NV_{min}^2(Y_2') = \frac{2n(N - 1)}{NV} [1 - n/N + \sqrt{2(1 - \delta)}]. \quad (1.15)$$

Echantillonnage avec PPT en deux occasions

N.G.N. PRASAD et J.E. GRAHAM¹

RÉSUMÉ

La "méthode des groupes aléatoires" pour un échantillonnage avec probabilité proportionnelle à la taille (PPT) est étendue à un échantillonnage effectué en deux occasions. Nous utilisons de l'information sur une variable d'étude observée à la première occasion pour prélever la portion appariée de l'échantillon à la deuxième occasion. Nous examinons deux ensembles de données réels en vue d'une illustration numérique et d'une comparaison avec d'autres méthodes existantes.

MOTS CLÉS: Estimateur composite; comparaisons d'efficacité; méthode des groupes aléatoires; probabilité proportionnelle à la taille.

1. INTRODUCTION

Il est très fréquent, dans les enquêtes à passages répétés, qu'un plan de sondage avec remplacement partiel soit utilisé, en raison notamment de l'efficacité accrue de l'estimation qu'on espère ainsi obtenir, et aussi parce que cela permet de réduire le fardeau de réponse. Essentiellement, après chaque passage, une fraction des unités observées est retranchée de l'échantillon et remplacée par un nouveau sous-échantillon fraîchement prélevé dans la population. Cet ensemble d'unités non appariées est observé au passage suivant, en même temps que l'ensemble restant d'unités appariées. La littérature regorge d'analyses sur l'échantillonnage avec probabilités de sélection égales en deux occasions et sur les méthodes d'estimation connexes. Un cas particulièrement important survient lorsque les unités, à une occasion donnée, sont prélevées avec des probabilités de sélection inégales. Dans les travaux publiés jusqu'ici, on se sert de l'information recueillie à l'occasion précédente en vue d'améliorer l'estimateur habituel du total ou de la moyenne pour l'occasion courante, en utilisant une méthode d'estimation fondée sur les différences. Dans le présent article, nous présentons un plan visant un échantillonnage en deux occasions, et une méthode d'estimation connexe, qui se servent d'une information recueillie à l'occasion 1 (précédente) pour la sélection du sous-échantillon à observer à l'occasion 2 (courante). Par souci d'exhaustivité en même temps que de concision, nous n'examinons dans la présente section que les méthodes de sélection avec probabilités inégales aux deux occasions.

Considérons une population finie de N unités, désignées 1, 2, ..., N , et un échantillonnage en deux occasions: 1 (occasion précédente) et 2 (occasion courante). Soient y_{1i} et y_{2i} les valeurs d'une caractéristique y pour la i ème unité observée à la première et à la deuxième occasion respectivement. Soient X_1 et X_2 les totaux respectifs pour la population. Supposons qu'une mesure de taille x soit connue pour chacune des unités de la population.

où

$$Y_1 = \sum_{i \in s_1} y_{1i} / (n p_i) \quad (1.1)$$

$$Y_2 = \bar{Q} Y_{2n} + (1 - \bar{Q}) Y_{2m}, \quad (1.2)$$

$$Y_{2n} = \sum_{i \in s_2} y_{2i} / (n p_i), \quad (1.3)$$

$$Y_{2m} = \sum_{i \in s_1} y_{1i} / (n p_i) + \sum_{i \in s_2} (y_{2i} - y_{1i}) / (m p_i), \quad (1.4)$$

et \bar{Q} est un coefficient de pondération, $0 \leq \bar{Q} \leq 1$. Si l'on suppose que

$$Y_1 = \sum_{i=1}^N (y_{1i} / p_i) - X_1, \quad Y_2 = \sum_{i=1}^N (y_{2i} / p_i) - X_2, \quad (1.5)$$

Raj (1965) a examiné le plan d'échantillonnage avec PPT (probabilité proportionnelle à la taille) suivant: à la première occasion, un échantillon s de taille n est prélevé avec probabilités p_i proportionnelles aux valeurs x_i , $i = 1, 2, \dots, N$, et avec remplacement. À la deuxième occasion, un échantillon aléatoire simple s_1 de m unités est prélevé dans s sans remplacement et un échantillon indépendant s_2 de $n - m$ unités est prélevé avec PPT et avec remplacement dans l'ensemble de la population. Les totaux Y_1 et Y_2 sont alors respectivement estimés sans biais par:

1.1 Le plan de Des Raj

¹ N.G.N. Prasad, Associate Professor, Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta, Canada T6G 2G1; J.E. Graham, Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

Voici pourquoi. D'après (6) et (18), nous avons

$$EQM(\bar{z}_{RS}) - EQM(\bar{z}_{RP}) = \frac{M}{K} \sum_{i=1}^K N_i \bar{z}_{RP}^2 + \left(\frac{1}{N_i} - \frac{1}{N_i} \right) D_i^2 \left(\frac{K N_i}{M} - 1 \right).$$

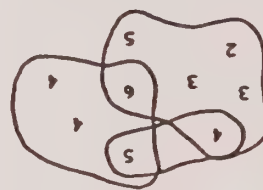
$$\times \left[(Z_i - \bar{Y} W_i)^2 + \left(\frac{1}{N_i} - \frac{1}{N_i} \right) D_i^2 \left(\frac{K N_i}{M} - 1 \right) \right].$$

Si la taille de la grappe, N_i , augmente, le facteur $(K N_i / M - 1)$ augmentera aussi. L'autre facteur du terme en sommation est $N_i [(Z_i - \bar{Y} W_i)^2 + (1/N_i - 1/N_i) D_i^2]$, il représente la part de l' EQM de \bar{z}_{RP} (équ. 18) attribuable à la variabilité de z et de w dans la grappe i (sans la constante $M/K N_i^2$). Si la taille de la grappe, N_i , augmente, la part de $EQM(\bar{z}_{RP})$ attribuable à la grappe i devrait aussi augmenter, ce qui fait que la covariance des deux facteurs est positive. Donc, l'estimateur \bar{z}_{RP} devrait avoir une EQM moins élevée que celle de \bar{z}_{RS} .

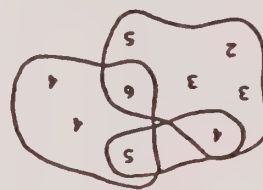
Exemple numérique. Dans cet exemple, nous appliquons les deux méthodes d'échantillonnage proposées à deux petites populations afin de mettre en lumière le calcul des valeurs F_{ij} , Z_{ij} et W_{ij} et la comparaison de ces valeurs. Pour les deux populations, $K = 3$, $k = 2$, $M = 12$ et $N = 9$. Il y a chevauchement des qu'une unité est présente dans plus d'une grappe à la fois. Les populations sont décrites dans le tableau 1.

Tableau 1
Comparaison des deux méthodes pour deux petites populations

N_i	n_i	Y_{ij}	F_{ij}	Z_{ij}	W_{ij}	$EQM(\bar{z}_{RS})$				$EQM(\bar{z}_{RP})$				$E.R.$	$B.R.(\bar{z}_{RS})$	$B.R.(\bar{z}_{RP})$
						3	4	5	2	1	4	5	2			
1	3	3,5,6	3,1,2	1,3,1,1	1,3,2,1,1	2,3,6,8,9	4,5	4,4,5,6	2,3,3,4,5,6	1,1,2,2	1,1,1,2,1,2	2,2	2,2	2	2	2
2	3	1,3,4,7	1,5,3	1,1,4,7	2,1,3,8,9	2,2,5	4,4,2,5,3	1,1,1,1/2,1/2	2,3,3,2,5,3	1,1,1,1/2,1/2	1,1,1,1/2,1/2	2,4	2,4	136,36	0,33	0,45
3	3	10,16	1,3,8	1,1/2,1/2,1/2	1,1/2,1/2,1/2	18,12	2,94	2,94	2,94	2,94	2,94	2,94	2,94	0,037	0,037	0,037



Population n° 1



Population n° 2

BIBLIOGRAPHIE

- AGARWAL, D.K., et SINGH, P. (1982). On cluster sampling strategies using ancillary information. *Sankhyā*, B, 44, 184-192.
- AMDEKAR, S.J. (1985). An unbiased estimator in overlapping clusters. *Bulletin of the Calcutta Statistical Association*, 15, 231-232.
- GIFFARD-JONES, W. (1993). The doctor game. *The Windsor Star*, April 15, 1993.
- GOEL, B.B.P.S., et SINGH, D. (1977). On the formation of clusters. *Journal of the Indian Society of Agricultural Statistics*, 29, 53-68.
- HANSEN, M.H., et HURWITZ, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- SÄRANDAL, C-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, S. (1988). Estimation in overlapping clusters. *Communications in Statistics, Theory and Methods*, 17, 613-621.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à une subvention (n° A-3111) du Conseil de recherches en sciences naturelles et en génie du Canada. Nous tenons aussi à exprimer toute notre reconnaissance aux arbitres et au rédacteur en chef, qui, par les précieux commentaires qu'ils ont faits sur une version antérieure de l'article, ont contribué à améliorer le produit.

4. CONCLUSION

Dans cet article, nous avons pu examiner la question du chevauchement des grappes sans nous embarrasser de l'hypothèse de la taille de population connue posée par Singh (1988). En outre, nous avons pu comparer les deux méthodes de façon plus directe, alors que Singh devait s'appuyer sur les résultats de Hansen et Hurwitz (1943).

L'analyse des résultats du tableau 1 confirme la théorie exposée dans cet article. Dans le cas des deux populations, le facteur $F = N_i [(Z_i - \bar{Y} W_i)^2 + (1/n_i - 1/N_i) D_i^2]$ augmente avec N_i , ce qui fait que $EQM(\bar{y}_{RP}) < EQM(\bar{y}_{RS})$, comme nous le disions plus haut.

$$BR(\bar{z}_{RP}) = \frac{M^2}{N^2} \left[\left(\sigma_{bz'w'}^2 - \frac{N^2}{YN} \right) \right]$$

$$+ \sum_{i=1}^I \frac{M}{N_i} \left(\frac{1}{1} - \frac{N_i}{1} \right) \left(S_{izw'}^2 - \frac{N^2}{YN} \right) \quad (14)$$

où

$$\sigma_{bz'w'}^2 = \sum_{i=1}^I \left(\bar{z}_i - Y/M \right) \left(\bar{w}_i - N/M \right) (N_i/M)$$

et $\sigma_{bz'w'}^2$ correspond à $\sigma_{bz'w'}^2$, z étant remplacé par w et Y , par N .

Démonstration. D'après un résultat courant, le biais relatif approximatif (au premier degré d'approximation) est

$$BR(\bar{z}_{RP}) = [V(N^{RP})/N^2]$$

$$- \text{Cov}(Y_{RP}, N_{RP})/YN. \quad (15)$$

Nous avons

$$V(N^{RP}) = V_1 E_2(N^{RP}) + E_1 V_2(N^{RP})$$

$$= M^2 \left[V_1 \frac{1}{k} \sum_{i=1}^I \bar{w}_i + E_1 \frac{1}{k} \sum_{i=1}^I \left(\frac{1}{1} - \frac{N_i}{1} \right) S_{izw'}^2 \right]$$

$$= \frac{M^2}{N^2} \left[\sigma_{bz'w'}^2 + \sum_{i=1}^I \frac{M}{N_i} \left(\frac{1}{1} - \frac{N_i}{1} \right) S_{izw'}^2 \right]. \quad (16)$$

De la même manière, nous pouvons écrire

$$\text{Cov}(Y_{RP}, N_{RP}) = \frac{M^2}{N^2} \left[\sigma_{bz'w'} + \sum_{i=1}^I \frac{M}{N_i} \left(\frac{1}{1} - \frac{N_i}{1} \right) S_{izw'} \right]. \quad (17)$$

En substituant (16) et (17) dans (15), nous obtenons l'équation (14).

Théorème 5. L'EQM de l'estimateur \bar{z}_{RP} (au premier degré d'approximation) est définie

$$\widehat{EQM}(\bar{z}_{RP}) = \frac{M}{N^2} \sum_{i=1}^I N_i$$

$$\times \left[(\bar{z}_i - Y/M)^2 + \left(\frac{1}{1} - \frac{N_i}{1} \right) D_{iz}^2 \right]. \quad (18)$$

Démonstration. Au premier degré d'approximation, nous écrivons

$$\widehat{EQM}(\bar{z}_{RP}) = [V(Y_{RP}) - 2Y \text{Cov}(Y_{RP}, N_{RP})]$$

$$+ Y^2 V(N^{RP}) / N^2. \quad (19)$$

En outre, d'après le théorème 2.5 de Singh (1988), nous avons, par analogie,

$$V(Y_{RP}) = \frac{M^2}{N^2} \sigma_{bz'}^2 + \frac{M^2}{N^2} \sum_{i=1}^I \frac{M}{N_i} \left(\frac{1}{1} - \frac{N_i}{1} \right) S_{iz}^2, \quad (20)$$

où $\sigma_{bz'}^2 = \sum_{i=1}^I (N_i/M) (\bar{z}_i - Y/M)^2$. En substituant (16), (17) et (20) dans l'équation (19), puis en simplifiant, nous obtenons l'équation (18).

Théorème 6. Un estimateur convergent de $\widehat{EQM}(\bar{z}_{RP})$ (au premier degré d'approximation) est défini

$$\widehat{EQM}(\bar{z}_{RP}) = \frac{M^2}{N_{RP}^2} \cdot \frac{k(k-1)}{1} \sum_{i=1}^I (\bar{z}_i - \bar{z}_{RP} \bar{w}_i)^2. \quad (21)$$

Démonstration. Comme les unités de sondage du premier degré sont prélevées suivant un EARPPT, l'argument précédent dans la démonstration du théorème 3 vaut pour cette démonstration-ci.

D'après l'équation (20) et les résultats 2.9.1 et 4.5.1 de Särndal, Swensson et Wretman (1992), nous pouvons définir un estimateur non biaisé de

$$\sigma_{bz'}^2 + \sum_{i=1}^I \frac{M}{N_i} \left(\frac{1}{1} - \frac{N_i}{1} \right) S_{iz}^2$$

par l'expression

$$s_{bz'}^2 = \frac{1}{k} \sum_{i=1}^I \left(\bar{z}_i - \bar{z}_i/k \right)^2. \quad (22)$$

De la même manière, si nous définissons $s_{bz'w'}$ et $s_{bz'}^2$, nous pouvons montrer que

$$\widehat{EQM}(\bar{z}_{RP}) = \frac{M^2}{N_{RP}^2} (s_{bz'}^2 - 2\bar{z}_{RP} s_{bz'w'} + \bar{z}_{RP}^2 s_{bz'w'}^2),$$

qui peut s'écrire comme en (21).

3. COMPARAISON DE L'EFFICACITÉ

Dans cette section, nous comparons l'efficacité des estimateurs utilisés respectivement dans les méthodes A et B. **Remarque.** L'estimateur \bar{z}_{RS} , de la méthode B, devrait normalement être plus efficace que l'estimateur \bar{z}_{RP} , de la méthode A. Voici pourquoi. D'après (6) et (18), nous avons

$$\widehat{EQM}(\bar{z}_{RS}) = \frac{K}{K} \sum_{i=1}^K N_i^2 \left[(Z_i - YW)_2 + \left(\frac{1}{1} - \frac{1}{N_i} \right) D_i^2 \right] \quad (6)$$

où $D_i^2 = S_{iz}^2 - 2Y S_{izw} + Y^2 S_{iww}^2$, et $S_{iz}^2 = \sum_{j=i}^{N_i} (Z_{ij} - \bar{Z}_i)^2 / (N_i - 1)$.

Démonstration. Au premier degré d'approximation, nous

avons

$$\widehat{EQM}(\bar{z}_{RS}) = [V(Y_{RS}) - 2Y \text{Cov}(Y_{RS}, N_{RS}) + Y^2 V(N_{RS})] / N^2. \quad (7)$$

D'après (4), on peut exprimer $V(Y_{RS})$ par la formule

$$V(Y_{RS}) = \frac{K}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{1}{N_i} \right) S_{iz}^2 \quad (8)$$

où $\sigma_{bz}^2 = \sum_{i=1}^K (N_i \bar{Z}_i - Y/K)^2 / K$.

En substituant (4), (5) et (8) dans l'équation (7), nous obtenons, après simplification,

$$\widehat{EQM}(\bar{z}_{RS}) = \frac{K}{K} \sum_{i=1}^K N_i^2 (\sigma_{bz}^2 - 2Y \sigma_{bz w} + Y^2 \sigma_{bw}^2) \quad (9)$$

En substituant dans (9) les expressions pour σ_{bz}^2 , $\sigma_{bz w}$ et σ_{bw}^2 , et en simplifiant, nous obtenons l'équation (6). Dans le théorème suivant, nous définissons un estimateur

de $\widehat{EQM}(\bar{z}_{RS})$.

Théorème 3. Un estimateur convergent de $\widehat{EQM}(\bar{z}_{RS})$ (au premier degré d'approximation) est défini

$$\widehat{EQM}(\bar{z}_{RS}) = \frac{K}{K} \sum_{i=1}^K N_i^2 (\bar{z}_i - \bar{z}_{RS} w_i)^2. \quad (10)$$

Démonstration. Nous remarquons que le premier degré d'échantillonnage consiste en un EASAR et que les variables aléatoires $N_i \bar{z}_i$ et $N_i w_i$, contenues dans l'estimateur par quotient, sont distribuées de façon indépendante mais identique. On peut donc estimer l'erreur quadratique moyenne de \bar{z}_{RS} en se fondant sur le résultat bien connu selon lequel l'estimateur de la variance dans un plan à plusieurs degrés peut ne prendre en compte que le premier degré (voir Särndal, Swensson et Wretman, 1992, résultats 2.9.1 et 4.5.1).

D'après (9), un estimateur sans biais de

$$\sigma_{bz}^2 + \frac{1}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{1}{N_i} \right) S_{iz}^2 \quad (11)$$

peut être défini

$$s_{bz}^2 = \frac{1}{K} \sum_{i=1}^K (N_i \bar{z}_i - \sum_{i=1}^K N_i \bar{z}_i / K)^2, \quad (12)$$

et un estimateur sans biais de

$$\sigma_{bz w} + \frac{1}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{1}{N_i} \right) S_{iz w}$$

est

$$s_{bz w} = \frac{1}{K} \sum_{i=1}^K (N_i \bar{z}_i - \sum_{i=1}^K N_i \bar{z}_i / K) \left(N_i w_i - \sum_{i=1}^K N_i w_i / K \right) \quad (13)$$

De la même manière, un estimateur indépendant de

$$\sigma_{bw}^2 + \frac{1}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{1}{N_i} \right) S_{iw}^2$$

est s_{bw}^2 , défini comme en (11).

A l'aide de ces résultats, il est facile de montrer qu'un estimateur convergent de $\widehat{EQM}(\bar{z}_{RS})$ (équ. 6) est défini par l'expression

$$\widehat{EQM}(\bar{z}_{RS}) = \frac{K}{K} \sum_{i=1}^K N_i^2 (s_{bz}^2 - 2s_{bz w} s_{bw} + s_{bw}^2)$$

qui peut s'écrire comme en (10).

2.2 Méthode B

Cette méthode comporte les étapes suivantes:

- Prélever k grappes parmi K au moyen d'un échantillonnage avec remise avec probabilité proportionnelle à la taille (EARPPT), les probabilités de sélection étant $P_i = N_i / M$, $i = 1, \dots, K$.

- Même étape que celle de la méthode A.

Théorème 4. L'estimateur par quotient suivant un échantillonnage PPT,

a pour biais relatif (au premier degré d'approximation)

Lorsqu'on dispose de données informatiques par grappe sur les unités, il est facile de connaître la valeur des F_{ij} . En reprenant l'exemple en épidémiologie cité plus haut, supposons qu'il existe des données informatiques sur des ménages ou des particuliers, accompagnées de codes d'identification pour ces unités (par ex., numéro civique, numéro d'assurance sociale ou numéro d'assurance-maladie). Alors, par une simple instruction de programme, un expérimentateur peut savoir facilement, grâce au code d'identification, combien de fois une certaine unité se répète dans les différentes grappes. En outre, si nous avons un diagramme des grappes chevauchantes et si le critère de formation des grappes ne permet pas d'éliminer les doubles, il est possible de connaître le nombre de fois que des unités sont observées dans les différentes grappes.

Nous examinons les deux méthodes d'échantillonnage dans la section 2 et nous comparons leur efficacité dans la section 3.

2. DESCRIPTION DES DEUX MÉTHODES

Les deux méthodes proposées sont décrites dans les sections 2.1 et 2.2. Nous en faisons la comparaison dans la section 3.

2.1 Méthode A

Cette méthode comporte les étapes suivantes:

- Prélever k grappes parmi K au moyen d'un échantillonnage aléatoire simple avec remise (EASAR).

- Dans la grappe échantillonnée i de taille N_i ($i = 1, \dots, K$), prélever n_i unités élémentaires par échantillonnage aléatoire simple sans remise (EASSR).

Théorème 1. L'estimateur par quotient suivant un EAS

$$\bar{z}_{RS} = Y_{RS}/N_{RS} = \frac{k}{K} \sum_{i=1}^K \frac{N_i \bar{z}_i}{N_i w_i} \bigg/ \frac{k}{K} \sum_{i=1}^K \frac{1}{N_i w_i} \quad (1)$$

a pour biais relatif (au premier degré d'approximation)

$$BR(\bar{z}_{RS}) \doteq \frac{k}{K} \left[\left(\frac{\sigma_{bw}^2}{N^2} - \frac{\sigma_{bzw}^2}{NY} \right) K \right]$$

$$+ \sum_{i=1}^K \frac{k}{K} N_i^2 \left(\frac{1}{1} - \frac{n_i}{N_i} \right) \left(S_{lw}^2 - S_{lzw}^2 \right) \quad (2)$$

où

$$\sigma_{bzw}^2 = \sum_{i=1}^K (N_i \bar{z}_i - Y/K) (N_i w_i - N/K) / K$$

$$S_{lzw}^2 = \sum_{i=1}^{N_i} (Z_{ij} - \bar{Z}_i) (W_{ij} - \bar{W}_i) / (N_i - 1),$$

et σ_{bw}^2 , S_{lw}^2 , \bar{W}_i et \bar{w}_i correspondent respectivement à σ_{bzw}^2 , S_{lzw}^2 , \bar{Z}_i et \bar{z}_i , z étant remplacé par w et Y , par N .

Démonstration. D'après un résultat courant, le biais relatif de l'estimateur \bar{z}_{RS} (au premier degré d'approximation) est

$$BR(\bar{z}_{RS}) \doteq [V(N_{RS})/N^2] - \text{Cov}(Y_{RS}, N_{RS})/YN. \quad (3)$$

Posons E_2 et V_2 comme l'espérance et la variance conditionnelles relatives à un échantillon donné de grappes, et E_1 et V_1 comme l'espérance et la variance relatives à tous les échantillons de grappes. Nous avons alors

$$V(N_{RS}) = V_1 E_2(N_{RS}) + E_1 V_2(N_{RS})$$

$$= V_1 \left[\frac{k}{K} \sum_{i=1}^K N_i E_2(w_i) \right]$$

$$+ E_1 \left[K^2 \sum_{i=1}^K N_i^2 V_2(w_i) \right]$$

$$= V_1 \left(\frac{k}{K} \sum_{i=1}^K N_i w_i \right)$$

$$+ E_1 \left[K^2 \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{n_i}{N_i} \right) S_{lw}^2 \right]$$

$$= \frac{k}{K^2} \sigma_{bw}^2 + \frac{k}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{n_i}{N_i} \right) S_{lw}^2. \quad (4)$$

De la même manière, nous avons

$$\text{Cov}(Y_{RS}, N_{RS}) = \frac{k}{K^2} \sigma_{bzw}^2$$

$$+ \frac{k}{K} \sum_{i=1}^K N_i^2 \left(\frac{1}{1} - \frac{n_i}{N_i} \right) S_{lzw}^2. \quad (5)$$

En substituant (4) et (5) dans (3), nous obtenons l'équation (2), ce qui complète la démonstration du théorème.

Théorème 2. L'erreur quadratique moyenne (EQM) de l'estimateur \bar{z}_{RS} (au premier degré d'approximation) est

Estimation pour grappes chevauchantes lorsque la taille de la population est inconnue

D.S. TRACY et S.S. OSAHAN¹

RÉSUMÉ

Singh (1988) propose deux méthodes d'échantillonnage en vue d'estimer la moyenne d'une population en grappes chevauchantes lorsque la taille de la population est connue. Dans cet article, nous étudions des estimateurs par quotient appliqués dans ces deux méthodes en supposant que la taille réelle de la population est inconnue, ce qui est plus conforme à la réalité des enquêtes par sondage. Nous comparons l'efficacité des estimateurs appliqués dans l'une et l'autre méthodes et nous donnons un exemple numérique.

MOTS CLÉS: Grappes chevauchantes; formation de grappes avant échantillonnage; erreur quadratique moyenne; efficacité relative.

1. INTRODUCTION

Dans l'échantillonnage par grappes, les grappes sont formées soit avant l'échantillonnage (FGAVE), soit après l'échantillonnage (FGAPE). Dans les deux cas, les grappes peuvent se chevaucher ou non. Il existe déjà beaucoup d'ouvrages de recherche sur les grappes non chevauchantes. Cependant, de nombreux cas d'échantillonnage impliquent des grappes chevauchantes. Par exemple, on peut retrouver de telles grappes dans une enquête épidémiologique régionale portant sur une maladie contagieuse comme la tuberculose, qui tend à se répandre avec la propagation du SIDA (Gifford-Jones 1993). Dans ce cas-ci, les grappes peuvent être formées de sujets infectés ou de personnes qui sont en rapport étroit avec eux et qui sont plus vulnérables à cette maladie. De même, dans une étude écologique, les grappes peuvent être formées de centrales au charbon qui émettent des hydrocarbures aromatiques polycycliques (HAP), composés doués de propriétés cancérogènes. Les grappes sont constituées en fonction de la concentration de ces gaz, et des études peuvent s'imposer dans le but de surveiller la pollution atmosphérique, qui est à l'origine d'infections pulmonaires telles la bronchite. Au sujet des grappes chevauchantes, on peut consulter les travaux effectués par Goel et Singh (1977), Agarwal et Singh (1982) et Amdekar (1985) sur certains aspects de la question. Cependant, les méthodes élaborées par ces auteurs présentent toutes des lacunes.

Il y a quelques années, Singh (1988) a élaboré un estimateur très simple pour la moyenne d'une population. Il utilise cet estimateur dans deux méthodes d'échantillonnage par grappes (selon la formule FGAVE) en supposant que la taille de la population est connue. La première méthode prévoit un échantillonnage des grappes avec probabilités égales tandis que la seconde prévoit un échantillonnage avec probabilités proportionnelles à la taille de la grappe. En ce qui concerne l'échantillonnage des éléments

des grappes, la règle des probabilités égales s'applique pour les deux méthodes. Toutefois, supposer que la taille de la population est connue n'est pas réaliste. Si c'était vraiment le cas, on connaîtrait *a priori* tous les éléments en double de la population et il suffirait de les éliminer pour accroître l'efficacité du plan d'échantillonnage. C'est pourquoi il est nécessaire d'améliorer les estimateurs de la moyenne d'une population de Singh (1988) afin de les rendre plus utiles, car ils dépendent de la taille réelle de la population. C'est exactement le but du présent article. Nous proposons deux méthodes d'échantillonnage par grappes (formule FGAVE) qui utilisent des estimateurs par quotient ordinaires de la moyenne de population; ces estimateurs ne dépendent pas de la taille réelle de la population. Comme dans Singh (1988), la première méthode consiste en un échantillonnage avec remise avec probabilités égales, tandis que la seconde consiste en un échantillonnage avec probabilités inégales. Dans les deux cas, les éléments des grappes sont prélevés suivant un plan d'échantillonnage sans remise avec probabilités égales. La population de N unités à l'étude peut être définie comme un ensemble de K grappes chevauchantes où N_i désigne le nombre d'unités contenues dans la grappe i et $\sum_{i=1}^K N_i = M \geq N$, la taille réelle (inconnue) de la population, (l'égalité ne vaut que dans le cas des grappes non chevauchantes). Une unité peut appartenir à plus d'une grappe à la fois. Posons y comme la caractéristique étudiée et \bar{Y} , comme la moyenne de la population.

Définissons

$$Z_{ij} = Y_{ij}/F_{ij}, \quad W_{ij} = 1/F_{ij}; \quad i = 1, 2, \dots, K, \quad j = 1, 2, \dots, N_i$$

où Y_{ij} est la valeur de y pour l'unité j dans la grappe i et F_{ij} est le nombre de fois que cette unité est observée dans les K grappes.

¹ D.S. Tracy et S.S. Oshan, Department of Mathematics and Statistics, University of Windsor, Windsor (Ontario), N9B 3P4.

complexes, l'effet de l'ajout de résidus théoriques aux données imputées pourrait, par exemple, être étudié. Toutefois, cette technique a comme unique objet la sous-estimation de V_{imp}^{sam} par V_{naive} , et ne tient pas compte de l'effet de V_{imp} . Enfin, d'autres paramètres, par exemple la médiane, et l'effet de l'imputation sur leur variance restent à étudier. Des extensions multidimensionnelles pourraient aussi être envisagées: l'estimation des corrélations, des quotients et des paramètres de régression en présence d'une imputation serait sans doute un sujet intéressant.

REMERCIEMENTS

Les auteurs tiennent à remercier M. J.N.K. Rao pour son appui et ses encouragements constants, ainsi que le rédacteur associé pour ses commentaires utiles.

BIBLIOGRAPHIE

- HANSEN, M., HURWITZ, W., et MADOW, W. (1953). *Sample Survey Methods and Theory*. (Volume 2), New York: J. Wiley, 139-141.
- KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

- LEE, H., RANCOURT, E., et SÄRNDAAL, C.-E. (1991). Experiments with variance estimation from survey data with imputed values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 690-695.
- RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Rapport non publié, Statistique Canada.
- RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Rapport non publié, Statistique Canada.
- RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- RANCOURT, E., LEE, H., et SÄRNDAAL, C.-E. (1992). Bias corrections for survey estimates from data with imputed values for nonignorable nonresponse. *Proceedings 1992 Annual Research Conference*, Bureau of the Census, 523-539.
- RANCOURT, E., LEE, H., et SÄRNDAAL, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 374-379.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.
- SÄRNDAAL, C.-E. (1990). Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation. Conférence spéciale invitée. *Recueil: Symposium 90, Mesure et amélioration de la qualité des données*, Statistique Canada, 337-350.

offrir, et le véritable intérêt réside dans l'estimation de l'erreur quadratique moyenne. Autrement dit, des efforts accrus doivent être axés sur l'amélioration des estimations ponctuelles et de leurs biais. Des résultats préliminaires dans ce domaine ont été présentés par Rancourt, Lee et Särndal (1992).

5. CONCLUSION

Il est bien connu que l'estimateur de la variance habituel sous-estime la variance de l'estimation de \bar{Y} en présence de valeurs imputées, si ces valeurs sont traitées comme des valeurs observées. Dans la présente étude, nous avons de

produit l'estimateur de la variance simpliste lorsque des données sont imputées. Nous avons examiné plusieurs méthodes d'imputation pour évaluer dans quelle mesure le degré de sous-estimation était lié à la méthode d'imputation. Nous avons évalué un estimateur de la variance jackknife unifié de la forme proposée par Rao et Shao (1992), lequel tient compte de la variance due au processus d'imputation. L'étude a révélé certaines propriétés intéressantes de l'estimateur proposé dans le cas aussi bien de l'échantillonnage aléatoire simple que de plans d'enquête complexes. Nos observations sont résumées dans les paragraphes qui suivent.

- (1) L'ampleur de la sous-estimation de la variance est fortement liée à la fois à la capacité de la méthode d'imputation de prédire les valeurs vraies, et à sa capacité de préserver la variation naturelle des données.
- (2) L'estimateur de la variance jackknife rajusté qui est proposé offre une approche unifiée pour l'estimation de la variance de données imputées, qui peut être appliquée facilement à un certain nombre de méthodes d'imputation et à des plans de complexité variable.
- (3) D'un point de vue pratique, aucune modification du fichier imputé initial n'est nécessaire, de sorte que l'estimation des moyennes et des totaux n'est aucunement perturbée par la nécessité d'estimer les variances.
- (4) La méthode proposée peut facilement être étendue à des plans plus complexes, à plusieurs classes d'imputation et, avec prudence, au cas d'une non-réponse non aléatoire qui dépend seulement des variables auxiliaires disponibles.
- (5) L'estimateur de la variance jackknife rajusté donne de bons résultats lorsque la non-réponse est uniforme ou que le modèle linéaire habituel est vérifié, ce qui découle du fait que l'estimateur est à la fois convergent selon le plan et non biaisé selon le plan et le modèle.
- (6) Dans le cas du modèle P_L , dans lequel les unités ayant des valeurs y élevées sont plus susceptibles d'être non répondantes, les trois estimateurs de la variance ont un rendement extrêmement faible.

- (7) Dans le cas de la non-réponse liée à la variable y , de meilleures techniques d'imputation sont nécessaires, et les estimateurs ponctuels doivent être étudiés plus à fond. Ici, il faut plutôt s'intéresser à l'estimation de l'erreur quadratique moyenne qu'à celle de la variance.

Les enquêtes actuelles étant soumises à un degré élevé d'imputation, du moins dans certaines classes d'imputation, il est clair qu'on ne peut passer sous silence l'effet de l'imputation sur l'estimation de la variance. Une surestimation de la précision peut produire des intervalles de confiance trop étroits et amener à présenter comme significatives des données qui ne le sont pas. Si la mise en oeuvre des méthodes suggérées ci-dessus est jugée trop coûteuse dans une situation particulière, il faudrait à tout le moins effectuer des études pour évaluer l'impact de l'imputation dans certains cas représentatifs. Un facteur spécial d'accroissement de la variance pourrait alors être appliqué. Toutefois, avec l'émergence de logiciels généraux d'estimation, il semble de moins en moins justifiable de ne pas recourir à des estimateurs de la variance qui font une juste évaluation de l'effet de l'imputation.

De nombreux problèmes demeurent évidemment non résolus, et sont peut-être impossibles à résoudre. Premièrement, l'imputation du plus proche voisin doit faire l'objet de travaux théoriques beaucoup plus intenses. Les rajustements de la méthode du jackknife que nous avons appliqués à cette méthode d'imputation ont donné un rendement moindre que ceux qui ont été appliqués aux autres méthodes. Il faudra peut-être trouver des fonctions plus lisses pour remplacer la méthode du plus proche voisin. Deuxièmement, la robustesse de l'estimateur proposé doit être évaluée. Il est clair qu'un rendement satisfaisant peut être obtenu si le modèle (15) est vérifié et que la non-réponse est aléatoire. Le non-respect partiel de l'une ou l'autre de ces conditions n'a pas semblé entacher de bon rendement de l'estimateur jackknife dans notre expérience limitée, mais d'autres travaux devront être entrepris dans cette voie. La dérogation aux deux conditions simultanément n'a pas encore été étudiée. Les cas de non-réponse non aléatoire dans lesquels la propension à ne pas répondre est liée à la variable y sont encore moins bien compris, bien qu'il faille mettre l'accent, dans ce cas, sur l'estimation de l'erreur quadratique moyenne plutôt que de la variance. Troisièmement, des comparaisons devraient être faites avec des résultats d'imputation multiple. Il faut reconnaître, toutefois, que des méthodes d'imputation appropriées (Rubin 1987) doivent d'abord être établies. Notons qu'aucune des méthodes d'imputation étudiées ici n'est pas appropriée en ce qui a trait à l'imputation multiple. Il faudrait étendre l'analyse à d'autres méthodes d'imputation et à d'autres paramètres d'intérêt. La présente étude s'est limitée à quatre méthodes d'imputation simples. En pratique, des méthodes beaucoup plus complexes sont utilisées, et elles sont souvent combinées les unes aux autres. L'impact sur l'estimation de la variance du recours à plus d'une méthode d'imputation a été étudié par Rancourt, Lee et Särndal (1993); d'autres travaux sont nécessaires. Pour ce qui est d'autres méthodes d'imputation plus

pourvu que le facteur de correction pour population finie soit omis et que $(n - 1)/n \equiv 1$ et $(m - 1)/m \equiv 1$. Les résultats sont résumés au tableau 6.

Tableau 6

Biais relatif de l'estimateur de la variance simpliste (v_{naive}), de l'estimateur de la variance jackknife rajusté et de l'estimateur de la variance de Särndal en vertu d'une non-réponse non aléatoire de 30%

Modèle de non-réponse	Estimateur de la variance	Méthode d'imputation	
		Quotient	PPV
P_L	v_{naive}	-22.7	-54.6
	v_J	3.9	-37.5
	v_S	-2.6	-36.8
	v_{naive}	-4.0	-0.7
P_S	v_J	3.7	7.2
	v_S	2.8	4.5
	v_{naive}		

Dans le cas de l'imputation par quotient, l'estimateur

de la variance simpliste donne des résultats très différents pour les deux modèles de non-réponse (-22.7% contre -4.0%). En effet, s'il est vrai que la réduction de la taille effective de l'échantillon tend à réduire la variance dans les deux cas, il y a sureprésentation des unités de grande taille parmi les unités manquantes dans le modèle P_L , ce qui tend à accentuer cet effet, alors que dans le modèle P_S , où il y a sureprésentation des petites unités parmi les unités manquantes, cet effet tend à être partiellement compensé. Deuxième observation, l'estimateur de la variance jackknife rajusté donne de bons résultats dans le cas de l'imputation par quotient, mais est plutôt déce-

vant dans le cas de l'imputation du plus proche voisin. Cela s'explique par le fait que le présent ensemble de données suit assez bien le modèle linéaire habituel (15) et que l'estimateur de la variance jackknife rajusté s'est révélé non biaisé selon le modèle (Rao 1992) pour l'imputation par quotient. Par ailleurs, le rajustement de la méthode du plus proche voisin quand la non-réponse n'est pas uniforme, L'autre rajustement applicable à l'imputation du plus proche voisin, que nous avons décrit à la section 3, donne également de piètres résultats en termes absolus (valeurs non présentées ici), bien que les estimations soient toujours prudentes. Troisième observation, la performance de l'estimateur de Särndal, v_S , équivalant grosso modo à celle de l'estimateur jackknife rajusté, en vertu aussi bien de la méthode du quotient que de celle du plus proche voisin, et d'une non-réponse non aléatoire qui dépend

Lorsque le mécanisme de réponse n'est pas aléatoire, et que la propension à répondre est reliée à la variable touchée par la non-réponse (y), les estimateurs ponctuels sont eux-mêmes gravement biaisés selon les quatre méthodes d'imputation. L'estimation de la variance a donc peu à

$$P_L = 1 - \exp(-c_L x), \quad (22)$$

$$P_S = \exp(-c_S x), \quad (23)$$

où les constantes c_L et c_S sont choisies de façon à donner un taux de non-réponse probable de 30%. Dans le modèle P_L donné en (22), la non-réponse est en corrélation positive avec la variable x , ce qui signifie qu'une non-réponse est plus probable dans le cas des grandes unités (L). L'inverse est vrai pour le modèle P_S donné en (23), en vertu duquel ce sont les petites unités (S) qui sont le plus susceptibles de ne pas répondre. Les méthodes d'imputation qui ne tiennent pas compte de la variable x (moyenne et hot-deck) devraient normalement produire des estimateurs de \bar{Y} qui sous-estiment la moyenne vraie en vertu du modèle de non-réponse (22) et qui surestiment la moyenne vraie en vertu du modèle (23). Toutefois, les méthodes d'imputation qui utilisent la variable auxiliaire (quotient et plus proche voisin) devraient produire de meilleures estimations de la moyenne. Une simulation, dont les résultats sont présentés au tableau 5 ci-dessous, a permis de confirmer ces suppositions. Comme auparavant, nous avons utilisé 100,000 répétitions.

Tableau 5

Estimations de la moyenne \bar{Y} en pourcentage de la moyenne vraie lorsque la non-réponse n'est pas aléatoire et que le taux de non-réponse probable est de 30%

Modèle de non-réponse	Méthode d'imputation		
	Moyenne	HD	Quotient
en pourcentage			
P_L	60.4	60.4	94.7
P_S	132.7	132.7	102.0
			101.4
			93.5

De toute évidence, l'estimation de la variance n'offre aucun intérêt lorsque les estimateurs ponctuels eux-mêmes sont fortement biaisés, comme c'est le cas pour les méthodes de la moyenne et du hot-deck. Toutefois, dans le cas des méthodes du quotient et du plus proche voisin, pour lesquelles les estimateurs ponctuels sont supérieurs, nous avons examiné la performance de l'estimateur de la variance jackknife rajusté, ainsi que d'un estimateur proposé par Särndal (1990), qui peut s'écrire (Rao 1992):

$$v_S(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{y}_m}\right)^2 \frac{1}{m(m-1)} \sum_{i \in s_I} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2 + \left(\frac{\bar{y}}{\bar{y}_m}\right)^2 \frac{n^2(m-1)}{2m} \sum_{i \in s_I} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2 + \left(\frac{\bar{y}}{\bar{y}_m}\right)^2 \frac{n(n-1)}{1} \sum_{i \in s} (x_i - \bar{x})^2,$$

réelle. La taille du ménage, connue pour tous les ménages de l'échantillon, a été utilisée pour former deux classes de l'échantillon, soit les ménages à un seul membre et les ménages à plus d'un membre. L'hypothèse sous-jacente était que la propension à répondre est différente entre ces deux classes; on a supposé, en revanche, une probabilité de réponse uniforme à l'intérieur des classes d'imputation. Deux structures de non-réponse ont été évaluées. La première suppose une non-réponse uniforme de 5% dans la classe des ménages à un seul membre et une non-réponse uniforme de 10% dans la classe des ménages à plusieurs membres, tandis que la deuxième suppose des taux de non-réponse de 25% et de 30% respectivement pour ces deux classes. L'imputation hot-deck, les rajustements des valeurs imputées et les calculs des totaux rajustés donnés en (20), $Y_v^j(g_j)$, ont été effectués indépendamment dans chaque classe d'imputation, désignée par v . Les termes $Y_v^j(g_j)$ ont ensuite été additionnés pour les deux classes d'imputation, ce qui a donné $Y_v^j(g_j)$, dont on s'est servi pour obtenir l'estimation v_j , selon l'équation (21). Les résultats sont résumés au tableau 4.

Tableau 4

Biais relatif et stabilité relative (entre parenthèses) de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, dans le cas d'un échantillonage stratifié à plusieurs degrés et de deux classes d'imputation

Taux de non-réponse		Taux de non-réponse	
5% et 10%	25% et 30%	en pourcentage	
Estimateur de la variance		v_{naive}	v_j
		-16.7 (87)	-1.0 (103)
		-40.2 (84)	1.1 (127)

Comme le montre le tableau 4, l'estimateur de la variance jackknife rajusté v_j donne de bons résultats en vertu des deux structures de non-réponse. Ces résultats, jumelés à ceux du tableau 3, démontrent la convergence et la stabilité relativement bonne de l'estimateur de la variance jackknife rajusté, même dans le cas de taux de non-réponse élevés.

4.3 Non-réponse non aléatoire

Comme nous l'avons vu ci-dessus, l'estimateur de la variance jackknife rajusté donne de bons résultats quand la non-réponse est aléatoire à l'intérieur des classes d'imputation. Pour étudier sa robustesse à l'égard de l'hypothèse d'un mécanisme de réponse uniforme, nous avons utilisé l'ensemble de données décrit à la section 2, et produit une non-réponse de la manière décrite dans Lee, Ranncourt et Särndal (1991). Plus précisément, nous avons supposé que la probabilité de non-réponse était reliée à la variable x de deux façons distinctes:

$$v_j(Y_j) = \sum_L \frac{n_g}{n_g - 1} \sum_{j=1}^g n_g (Y_v^j(g_j) - Y_j)^2 \quad (21)$$

puis en posant

$$Y_v^j(g_j) = S(g_j) + \sum_{(hik) \in s-j} w_{hik} z_{hik}^{(g_j)} + \sum_{(hik) \in s-j, i \neq j} \frac{n_g}{n_g - 1} w_{gik} z_{gik}^{(g_j)} \quad (20)$$

Il peut être démontré que l'estimateur v_j défini en (21) est un estimateur convergent de la variance de Y_j (Rao et Shao 1992). Nous avons produit 10,000 échantillons de 60 upé choisies avec probabilité proportionnelle à la taille, et nous avons soumis les ménages sélectionnés à des taux de non-réponse uniformes de 5% et de 30%. Nous avons ensuite calculé l'estimateur de la variance simpliste, ainsi que l'estimateur de la variance jackknife rajusté, v_j , donné en (21). Le biais relatif (8) et la stabilité relative (9) ont été calculés pour les deux estimateurs de la variance, et les résultats sont résumés au tableau 3.

Tableau 3

Biais relatif et stabilité relative (entre parenthèses) de l'estimateur de la variance simpliste (v_{naive}) et de l'estimateur de la variance jackknife rajusté, dans le cas à des taux de non-réponse de 5% et de 30%, dans le cas d'un échantillonage stratifié à plusieurs degrés

Taux de non-réponse		Taux de non-réponse	
5%	30%	en pourcentage	
Estimateur de la variance		v_{naive}	v_j
		-10.3 (88)	-0.9 (97)
		-43.7 (84)	1.2 (124)

Comme en fait foi le tableau 3, l'estimateur de la variance simpliste sous-estime la variance vraie de Y en des proportions comparables à celles observées dans le cas de l'échantillonage aléatoire simple (tableau 2), et la sous-estimation s'aggrave avec l'augmentation du taux de non-réponse. L'estimateur de la variance jackknife rajusté, par contre, donne de bons résultats aux deux niveaux de non-réponse, au prix relativement modique d'une légère diminution de sa stabilité, en comparaison de v_{naive} .

4.2 Classes d'imputation

Selon le même plan d'échantillonnage que celui décrit à la section 4.1, nous avons aussi examiné le cas où il existe plus d'une classe d'imputation, qui reflète la situation

SUC

tre
suc
uon

Taux de non-réponse	Estimateur de la variance	Méthode d'imputation		
		Moyenne	HD	Quotient PPV
50%	V_{naive}	-10,7	-9,4	-2,5
	V_f	2,7	3,6	3,7
30%	V_{naive}	-51,4	-43,4	15,3
	V_f	3,3	1,9	3,0
				5,3

en pourcentage

de 50% et de 300% ont été utilisées et les biais relatifs ont été calculés. Les résultats sont résumés dans le tableau 2.

Puisque l'estimateur de la variance jackknife rajusté est convergent selon le plan (d -convergent) (Rao 1992), il est aussi convergent pour la variance jackknife rajustée. L'estimateur imputé de Y est alors donné par

$$Y_i = \sum_{(hik) \in \mathcal{S}_Y} w_{hik} y_{hik} + \sum_{(hik) \in \mathcal{S}_Y^*} w_{hik} y_{hik}^* \quad (16)$$

4. EXTENSIONS

$$S(g_j) = \sum_{h_{ik} \in \mathcal{H}} w_{h_{ik}} y_{h_{ik}} + \frac{n_g - 1}{n_g} \sum_{g_{ik} \in \mathcal{G}} w_{g_{ik}} y_{g_{ik}}, \quad (17)$$

4.1 Plans complexes

Dans la présente section, nous décrivons une étude de simulation qui évalue l'estimateur de la variance jackknife rajusté de Rao et Shao (1992) par rapport à l'estimateur de la variance simpliste, dans le cas d'un échantillonnage stratifié à plusieurs degrés et d'une imputation hot-deck. Plus précisément, nous utiliserons des données de l'enquête canadienne sur les finances des consommateurs (EFC), dont le plan de sondage est identique à celui de l'enquête sur la population active. La variable à l'étude, y , est le revenu total du ménage. L'EFC se fonde sur un plan complexe à plusieurs degrés avec stratification, et les unités primaires d'échantillonnage (upé) des strates utilisées dans la présente étude sont sélectionnées avec probabilité proportionnelle au nombre de logements. De façon générale, les upé sont des ensembles de logements, plus précisément des îlots urbains dans les villes et des groupes de secteurs

quand la (g_j) ième upé est supprimée:

Ensuite, comme dans les équations (12) et (13), on évalue l'estimateur de la variance jackknife en calculant d'abord, l'estimateur de la variance $z_{hik}^{(g_j)}$ et les unités y_{hik}^* +

$$z_{hik}^{(g_j)} = y_{hik}^* + \left[\frac{S(g_j)}{S} - \frac{T(g_j)}{T} \right]. \quad (19)$$

en posant

(g_j)ième upé est supprimée, (h_i) \neq (g_j), et (h_{ik}) \in $S - S_{g_j}$, lorsque la j ième upé de la g ième strate est supprimée. On effectue le rajustement des valeurs imputées quand la

$$T(g_j) = \sum_{\substack{(h_{ik}) \in S_{g_j} \\ h \neq g}} w_{hik} + \sum_{\substack{(g_{ik}) \in S_{g_j} \\ i \neq j}} \frac{n_{g_{ik}} - 1}{w_{g_{ik}}},$$

aléatoire simple, pour un mécanisme de non-réponse uni-forme et une seule classe d'imputation, nous examinons l'ackknfite "à une suppression" sont alors données par

$$E_*(Y_I) = \left[\sum_{(hik) \in s_I} w_{hik} y_{hik} / \sum_{(hik) \in s_I} w_{hik} \right] \times \sum_{(hik) \in s} w_{hik} \quad (17)$$

Shao 1992). L'espérance de Y_i en vertu de la méthode du hot-deck peut s'écrire ainsi (Rao et Shao 1992):

$$Y_I = \sum_{(hik) \in s_I} w_{hik} y_{hik} + \sum_{(hik) \in s_{-I}} w_{hik} y_{hik}^*, \quad (16)$$

	v_j
30%	-51.4
$v_j^{naïve}$	-43.4
	15.3
	3.0
	5.3

[illegible]

$$\begin{aligned} S(g_j) &= \sum_{\substack{(hik) \in \mathcal{C}_r \\ h \neq g}} w_{hik} y_{hik} + \frac{n_g - 1}{n_g} \sum_{\substack{(gik) \in \mathcal{C}_r \\ i \neq j}} w_{gik} y_{gik}, \\ T(g_j) &= \sum_{\substack{(hik) \in \mathcal{C}_r \\ h \neq g}} w_{hik} + \frac{n_g - 1}{n_g} \sum_{\substack{(gik) \in \mathcal{C}_r \\ i \neq j}} w_{gik}, \end{aligned} \quad (18)$$

stratifié à plusieurs degrés et d'une imputation hot-deck. Plus précisément, nous utiliserons des données de l'enquête canadienne sur les finances des consommateurs (EFC), dont le plan de sondage est identique à celui de l'enquête en posant

$$z_{(g)hik}^{*} = y_{hik}^{*} + \left[\frac{\hat{S}(g_j)}{\hat{S}} - \frac{\hat{T}(g_j)}{\hat{T}} \right] \quad (19)$$

la présente étude sont sélectionnées avec probabilité proportionnelle au nombre de logements. De façon générale, les upé sont des ensembles de logements, plus précisément des îlots urbains dans les villes et des groupes de secteurs quand la (g)/ième upé est supprimée:

3.1 Rajustement des valeurs imputées

Afin de produire la version "appropriée" (Rao 1990)

de l'estimateur de la variance jackknife, Rao (1992) a proposé de rajuster les valeurs imputées de la façon décrite ci-dessous. Intuitivement, un tel rajustement est nécessaire dès qu'une unité répondante est retirée d'une répétition jackknife, car pour la plupart des méthodes d'imputation, toutes les valeurs imputées dépendent directement ou indirectement de la valeur observée qui a été supprimée. Cela est clair dans le cas de l'imputation de la moyenne et de l'imputation par quotient, car tous les répondants contribuent directement à la moyenne \bar{y}_m , mais est moins évident pour les méthodes du plus proche voisin et du hot-deck, dans lesquelles l'unité supprimée contribue au processus d'imputation uniquement en ce sens qu'elle n'est pas disponible pour être choisie comme donneur. Ainsi, dès qu'une unité répondante est supprimée, toutes les valeurs imputées de l'échantillon doivent être rajustées avant que l'estimateur imputé "à une suppression" de la moyenne soit calculé. Le rajustement, de toute évidence, doit être fonction de la méthode d'imputation utilisée. Dans le cas des méthodes d'imputation de la moyenne et du hot-deck, on peut montrer que le rajustement suivant est approprié (Rao 1992; Rao et Shao 1992). Soit $z_i^*(j)$ la valeur rajustée de la i ème unité imputée y_i^* , quand la j ème unité a été supprimée. Alors, $z_i^*(j)$ est donnée par

$$z_i^*(j) = \begin{cases} y_i^* & \text{si } j \in s_r, \\ y_i^* + [\bar{y}_m(j) - \bar{y}_m] & \text{si } j \in s_p. \end{cases} \quad (11)$$

Autrement dit, aucun rajustement n'est nécessaire si l'unité supprimée (j) a elle-même été imputée, c'est-à-dire si l'unité j est un non-répondant. Dans le cas de l'imputation de la moyenne, par exemple, quand $j \in s_p$, la valeur rajustée se réduit à $\bar{y}_m(j)$, la moyenne des $m - 1$ répondants restants, comme souhaité.

On évalue l'estimateur de la variance jackknife d'abord en calculant l'estimateur imputé rajusté $y_j^d(j)$ suivant

$$y_j^d(j) = \sum_{\substack{i \in s \\ i \neq j}} z_i^*(j) / (n - 1), \quad (12)$$

puis en posant

$$v_j(y_j) = \frac{n}{n-1} \sum_{j=1}^n [y_j^d(j) - y_j]^2. \quad (13)$$

Il peut être démontré que l'estimateur de la variance jackknife rajusté se réduit à l'estimateur de la variance approprié dans le cas de l'imputation de la moyenne (Rao 1990) et offre une estimation convergente dans le cas de l'imputation hot-deck (Rao et Shao 1992). Pour ce qui est de l'imputation par quotient, les valeurs rajustées sont données par

3.2 Résultats empiriques

L'estimateur de la variance jackknife, avec rajustements correspondants aux quatre méthodes d'estimation décrites ci-dessus, a été calculé en sus de v_{naive} dans l'étude de simulation décrite à la section 2. Des taux de non-réponse

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i, \quad \text{cov}_m(y_i, y_j) = 0 \quad i \neq j \in s. \quad (15)$$

Puisque l'estimateur de la variance simpliste en vertu de l'imputation du plus proche voisin a été appliqué à l'imputation du plus proche voisin. Nous avons aussi examiné un autre rajustement, qui consistait à faire une nouvelle imputation de l'unité i au moyen de la méthode du plus proche voisin, lorsque l'unité supprimée (j) avait servi à l'imputation de l'unité i . Autrement dit, le rajustement a lieu seulement si l'unité supprimée est un répondant (comme ci-dessus), et seulement les non-répondants de la j ème répétition jackknife à qui l'unité j a été imputée font l'objet d'une nouvelle imputation pour recevoir la valeur de l'un des $m - 1$ donneurs restants. (Cela équivaut à l'imputation du deuxième voisin le plus proche pour ces unités.) Il convient de signaler qu'il n'existe pas de justification théorique pour l'un ou l'autre de ces rajustements. Puisque le dernier rajustement a donné un rendement inférieur à celui du rajustement de la méthode du quotient dans nos exemples, et qu'il serait lourd à appliquer en pratique, nous ne l'avons pas analysé davantage, bien qu'il ait toujours produit des estimations prudentes.

$$z_i^*(j) = \begin{cases} y_i^* & \text{si } j \in s_r, \\ y_i^* + \left[\bar{y}_m(j) \frac{x_i}{x_i - \bar{y}_m} x_i - \bar{y}_m \right] & \text{si } j \in s_p, \end{cases} \quad (14)$$

où $\bar{x}_m(j)$ est la moyenne des $m - 1$ valeurs de x pour les unités répondantes lorsque l'unité j est supprimée. L'estimateur de la variance jackknife $v_j(y_j)$ est alors calculé comme en (13) ci-dessus, ce qui donne l'estimateur de la variance approprié. En outre, Rao (1992) montre non seulement que l'estimateur de la variance jackknife rajusté est convergent selon le plan (p -convergent) dans le cas d'une non-réponse uniforme et indépendamment du modèle, mais aussi qu'il est non biaisé selon le plan et le modèle (p et m -non biaisé), en vertu du modèle (15) et de tout mécanisme de non-réponse qui ne dépend pas des valeurs y .

Tableau 1

(*naïve*) selon quatre méthodes d'imputation, à des taux de non-réponse de 5% et de 30%

Taux de	Estimateur	Méthode d'imputation	non- réponse	variance de la	Moyenne	HD	Quotient	PPV
5%	$V(f)$	9.9	10.3	9.5				
	v^{naive}	8.9	9.4	9.2				
	Biais rel. (v^{naive})	-10.7%	-9.4%	-2.5%				
30%	$V(f)$	13.5	16.5	10.1				
	v^{naive}	6.5	9.4	8.5				
	Biais rel. (v^{naive})	-51.4%	-43.4%	-15.3%				

voisin, puisque V_{imp} diminue avec la capacité de la méthode dans les cas des méthodes du quotient et du plus proche

(Sarnad 1990), comme c'est le cas dans la présente étude en raison de la corrélation relativement élevée entre les variables x et y . Un autre fait que révèle le tableau 1, c'est

de (\hat{p}_i) , quand les valeurs imputées sont traitées comme des valeurs observées, s'accroît la proportion de valeurs manquantes. Le problème est plus

auxiliaire. Signaux comme une sous-estimation de la variance

plus loin.

JACKKNIFE

3. ESTIMATEUR DE LA VARIANCE

plus loin.

plus proche voisin, caractéristique dont nous nous servons

comportement semblable des méthodes du quotient et du

le sont pas. Il est intéressant, par ailleurs, de noter le

amener à déclarer comme significatifs des résultats qui ne

des intervalles de confiance trop étroits d'environ 30% et

de l'ordre de 50% comme celle observée ici peut donner

auxiliaire. Signalons qu'une sous-estimation de la variance

moyenne et du hot-deck, qui n'utilisent pas d'information

prononcé dans le cas des méthodes d'imputation de la

la proportion de valeurs manquantes. Le problème est plus

des valeurs observées, s'amplifie à mesure que s'accroît

de $V(\bar{y})$, quand les valeurs imputées sont traitées comme

vation du taux de non-réponse. Ainsi, la sous-estimation

que $V(\bar{y})$ augmente, tandis que $V^{non-réponse}$ diminue, avec l'élé-

variables x et y . Un autre fait que révèle le tableau 1, c'est

en raison de la corrélation relativement élevée entre les

(Särndal 1990), comme c'est le cas dans la présente étude

d'imputation d'estimer les valeurs manquantes vraies

voisin, puisque $V^{non-réponse}$ diminue avec la capacité de la méthode

dans le cas des méthodes du quotient et du plus proche

plus grande). Par contre, $V(\bar{y})$ est légèrement plus faible

inhérente au hot-deck (c.-à-d. que la composante $V^{non-réponse}$ est

Taux de réponse non-estimateur	Méthode d'imputation	Moyenne	HD	Quotient	PPV
5%	$V(\hat{p})$	9.9	10.3	9.5	9.5
	v^{naive}	8.9	9.4	9.2	9.3
	Biais rel. (v^{naive})	-10.7%	-9.4%	-2.5%	-2.2%
30%	$V(\hat{p})$	13.5	16.5	10.1	10.3
	v^{naive}	6.5	9.4	8.5	9.0
	Biais rel. (v^{naive})	-51.4%	-43.4%	-15.3%	-12.8%

(*naïve*) selon quatre méthodes d'imputation, à des taux de non-réponse de 50% et de 30%

Soit $y_i(f)$ l'estimateur imputé de Y obtenu lorsque la i ème unité est retirée de l'échantillon. Dans le cas de l'échantillonnage aléatoire simple, un estimateur de la variance jackknife simplifiée de y_i est alors donné par

$$(10) \quad {}_2[I_{\underline{A}} - (f)I_{\underline{A}}] \sum_{u=1}^f \frac{u}{1-u} = f_{\underline{A}}$$

qui se réduit à V^{naive} , comme cela a été démontré (Rao 1992).

2. CONTEXTE

Suivant la notation de Rao (1992), nous supposons que dans un échantillon s de taille n , m unités répondent à la question y , tandis que $n - m$ unités ne le font pas. Désignons par y_i^* la valeur imputée pour l'unité i , $i \in s$, où s est l'ensemble des unités qui ont répondu. L'estimateur habituel de la moyenne \bar{y} basé sur le fichier soumis à l'imputation, dans le cas d'un échantillonnage aléatoire simple, est donné par

$$\bar{y}_I = \frac{1}{n} \left(\sum_{i \in s} y_i + \sum_{i \in s^c} y_i^* \right). \quad (1)$$

2.1 Méthodes d'imputation

Dans la présente étude de simulation, nous examinons quatre méthodes d'imputation simple: moyenne des répondants, quotient, plus proche voisin et hot-deck. Le lecteur est invité à lire l'article de Kalton et Kasprzyk (1986) pour un examen approfondi de la question de l'imputation. La méthode d'imputation la plus simple et la plus intuitive quand on veut estimer la moyenne des réponses à une question y consiste à imputer la moyenne des unités répondantes observées aux unités non répondantes. La valeur imputée y_i^* de l'unité i est donc la suivante, dans le cas de la méthode d'imputation de la moyenne:

$$y_i^* = \bar{y}_m = \sum_{j \in s_r} y_j / m. \quad (2)$$

L'estimateur de la moyenne de la population \bar{Y} donné en (1) se réduit alors à l'estimateur $\bar{y}_I = \bar{y}_m$. Comme cette méthode a l'inconvénient de causer une distorsion des distributions, elle n'est généralement utilisée qu'en dernier recours. Nous la présentons ici à des fins d'illustration. En second lieu, nous examinons une méthode d'imputation par quotient reposant sur l'hypothèse qu'une variable auxiliaire corrélée x est disponible et que le quotient y/x est le même dans les ensembles s_r et s^c , comme ce serait le cas si la non-réponse survenait au hasard, par exemple. Selon cette méthode du quotient, nous imputons la valeur suivante à la place de la valeur manquante y_i :

$$y_i^* = \frac{\bar{y}_m}{\bar{x}_m} x_i, \quad (3)$$

où \bar{x}_m est la moyenne des valeurs x de l'ensemble des répondants s_r . L'estimateur de la moyenne de la population \bar{Y} donné en (1) se réduit à l'estimateur d'échantillonnage double $\bar{y}_I = (\bar{y}_m / \bar{x}_m) \bar{x}$, si on considère les répondants comme l'échantillon de deuxième phase.

La troisième méthode d'imputation examinée est celle du plus proche voisin (PPV). Cette méthode consiste à imputer à une donnée manquante la valeur observée d'une autre unité de l'ensemble s_r , dont la distance par rapport

à l'unité non répondante est minimale. En pratique, les fonctions de distance utilisées sont habituellement les normes \hat{f}_1 , \hat{f}_2 , ou \hat{f}_∞ de Minkowski fondées sur les variables auxiliaires x , qu'on suppose observées pour toutes les unités de s . Ainsi

$$y_i^* = y_j, \quad j \in s_r, \quad \text{tel que } \|x_i - x_j\| \text{ est minimisé,} \quad (4)$$

où $\|\cdot\|$ désigne l'une des normes ci-dessus.

Les trois méthodes qui viennent d'être décrites sont souvent qualifiées de déterministes, car pour un échantillon de répondants donné, les valeurs imputées sont déterminées de façon unique. La quatrième méthode d'imputation examinée dans cette étude, la méthode du hot-deck (HD), est non déterministe, puisque les valeurs imputées sont choisies au hasard dans l'échantillon de répondants. Bien qu'en pratique des classes d'imputation soient souvent créées et qu'un processus séquentiel quelconque soit généralement mis en oeuvre, nous examinons ici le hot-deck pur, dans lequel le donneur (j) est choisi au hasard, avec remplacement, dans l'ensemble s_r complet, c.-à-d.

$$y_i^* = y_j, \quad j \in s_r. \quad (5)$$

2.2 Variance due à l'imputation

Si l'on traite les valeurs imputées comme des valeurs observées, on a l'estimateur incorrect de la variance

$$V^{naive} = (1 - f) s_I^2 / n, \quad (6)$$

où s_I^2 est la variance de l'échantillon complet (valeurs des répondants et valeurs imputées) et $(1 - f)$ est le facteur de correction pour population finie ($f = n/N$). Il est facile de montrer que la variance vraie de l'estimateur \bar{y}_I dans (1), $V(\bar{y}_I)$, peut s'écrire ainsi (Särndal 1990):

$$V(\bar{y}_I) = V^{sam} + V^{imp} + V^{mix}, \quad (7)$$

où V^{sam} est la variance d'échantillonnage, V^{imp} est la variance introduite par la méthode d'imputation en cause et V^{mix} est un terme de covariance entre V^{sam} et V^{imp} qui, dans la plupart des cas, est négligeable ou nul. On pourrait obtenir une estimation de V^{sam} en ajoutant à V^{naive} un terme de correction tenant compte du fait que la formule habituelle sous-estime la variance d'échantillonnage quand l'ensemble de données contient des valeurs imputées. Pour estimer $V(\bar{y}_I)$, toutefois, il est nécessaire d'estimer une composante de variance additionnelle, V^{imp} , attribuable au mécanisme d'imputation. Cela peut se faire explicitement, comme dans l'imputation multiple de Rubin (1987), ou encore en modifiant des formules de variance commune comme dans Särndal (1990) et dans Rao et Shao (1992). Notons que l'intérêt réside dans l'estimation de la variance de l'estimateur en cause, soit $V(\bar{y}_I)$, et non de la variance d'un estimateur qui aurait été obtenu en l'absence de non-réponse.

Méthode du jackknife pour l'estimation de la variance en présence de données imputées

J.G. KOVAR et E.J. CHEN¹

RÉSUMÉ

L'imputation est une méthode dont se servent couramment les organismes d'enquête afin de corriger le problème posé par la non-réponse à des questions particulières. Bien que dans la plupart des cas, les ensembles de données ainsi complétés offrent de bonnes estimations des moyennes et des totaux, les variances correspondantes, souvent sont largement sous-estimées. Plusieurs méthodes permettent de remédier à ce problème, mais la plupart dépendent du plan d'échantillonnage et de la méthode d'imputation. Récemment, Rao (1992) et Rao et Shao (1992) ont proposé une méthode jackknife unifiée pour l'estimation de la variance d'ensembles de données ayant fait l'objet d'une imputation. Le présent article évalue cette technique de manière empirique, au moyen d'une population réelle d'entreprises, et selon un plan d'échantillonnage aléatoire simple et un mécanisme de non-réponse uniforme. La possibilité d'étendre cette méthode à des plans d'échantillonnage stratifié à plusieurs degrés est examinée, et l'on se penche brièvement sur la performance de l'estimateur de la variance proposé dans le cas de mécanismes de réponse qui ne sont pas uniformes.

MOTS CLÉS : Non-réponse à des questions particulières; imputation hot-deck; imputation du plus proche voisin; non-réponse non aléatoire; plan d'échantillonnage complexe.

1. INTRODUCTION

Le problème de la non-réponse touche toutes les enquêtes par sondage, à des degrés divers. Dans le cas de la non-réponse totale d'une unité de l'échantillon, cette lacune est souvent corrigée par un rajustement approprié de la pondération de l'enquête, mais si la non-réponse ne touche que des questions particulières, la plupart des organismes d'enquête recourent à l'imputation. Ainsi, des valeurs plausibles sont insérées à la place de valeurs manquantes ou incohérentes, ce qui simplifie l'estimation des moyennes et des totaux à tous les niveaux d'agrégation. Dès les années 1950, toutefois, Hansen, Hurwitz et Madow (1953) ont reconnu que le fait de traiter les valeurs imputées comme des valeurs observées peut entraîner une sous-estimation des variances des estimateurs si les formules habituelles sont utilisées, sous-estimation qui s'amplifie à mesure que la proportion des réponses imputées s'accroît. Plusieurs solutions ont été suggérées pour résoudre ce problème. En particulier, Rubin (1987) a proposé de recourir à l'imputation multiple pour estimer la variance attribuable à l'imputation, en répétant le processus plusieurs fois et en estimant la variation entre les divers ensembles de données ainsi obtenus. Plus récemment, Särndal (1990) a présenté un certain nombre d'estimateurs de la variance fondés sur un modèle, tandis que Rao et Shao (1992) ont proposé une technique de rajustement des valeurs imputées qui permet de corriger l'estimateur de la variance jackknife habituel (ou simpliste). Les méthodes de Särndal et de Rao et Shao sont attrayantes du fait que seul le fichier

soumis à l'imputation (dans lequel les valeurs imputées sont marquées) est nécessaire à l'estimation de la variance. Aucun fichier auxiliaire n'est requis. La méthode de Särndal fondée sur un modèle donne des estimateurs sans biais de la variance pourvu que le modèle se vérifie (Lee, Rancourt et Särndal 1991). La méthode jackknife rajustée de Rao et Shao donne une estimation convergente selon le plan et non biaisée selon le modèle (Rao 1992). Toutefois, la méthode fondée sur un modèle exige des estimateurs de la variance différents pour chaque méthode d'imputation, tandis que la méthode jackknife rajustée offre une approche unifiée exigeant la mise en oeuvre d'un seul estimateur, l'estimateur jackknife, pourvu que les valeurs imputées soient convenablement rajustées à l'étape de l'estimation de la variance. Dans le présent article, nous décrivons une étude de simulation qui permet d'évaluer l'estimateur de la variance jackknife rajustée de Rao et Shao (1992). Dans la section 2, nous faisons valoir le bien-fondé de la présente étude empirique en montrant les caractéristiques de l'estimateur de la variance simpliste selon quatre méthodes d'imputation, dans le cas d'un échantillonnage aléatoire simple. Dans la section 3, nous décrivons brièvement la méthode de rajustement de Rao et Shao et nous présentons les résultats empiriques. Des extensions à des plans plus complexes et à des expériences comportant des mécanismes de non-réponse non aléatoires sont examinées à la section 4. Enfin, la section 5 contient des conclusions et des recommandations, et propose notamment des voies futures de recherche.

¹ J.G. Kovar, Division des méthodes d'enquêtes-entreprises; E.J. Chen, Division des méthodes d'enquêtes sociales, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa, Ontario, K1A 0T6.

- SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SETHI, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.
- SINGH, A.C., et MANTEL, H.J. (1991) Estimation composite par espace d'états pour les petites régions. *Recueil: Symposium 91: Conception et analyse des enquêtes longitudinales*, Statistique Canada, Ottawa, Novembre 1991, 17-25.
- TILLER, R. (1992). Time series modelling of sample data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 149-166.
- RAO, J.N.K., et YU, M. (1992). Small Area Estimation by Combining Time Series and Cross-sectional Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1-9.
- SALLAS, W.M., et HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.
- SÄRNDAAL, C.-E., et HIDIROGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHAIKLE, W.L. (1978). Choosing weights for composite estimation for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

BIBLIOGRAPHIE

- BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error-components model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- BELL, W.R., et HILLMER, S.C. (1987). Time series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.
- BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées modélisation et estimation. *Techniques d'enquête*, 15, 31-48.
- DUNCAN, D.B., et HORN, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association*, 67, 815-821.
- ERIKSEN, E.P. (1974). A regression method for estimating populations of local areas. *Journal of the American Statistical Association*, 69, 867-875.
- FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21, U.S. Office of Management and Budget.
- GHOSH, M., et RAO, J.N.K. (1994). Small Area Estimation: an Appraisal. *Statistical Science*, 9, à paraître.
- GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: University Press.
- JULIEN, C., et MARANDA, F. (1990). Sample design of the 1988 national farm survey. *Survey Methodology*, 16, 117-129.
- MANTTEL, H.J., SINGH, A.C., et BURBAU, M. (1993). Benchmarking of small area estimators. *Proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, 920-925.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economics Statistics*, 9, 163-175.
- PFEFFERMANN, D., et BARNARD, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics*, 9, 73-84.
- PFEFFERMANN, D., et BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- PLATEK, R., RAO, J.N.K., SÄRNDA, C.-E., et SINGH, M.P. Eds. (1987). *Small Area Statistics: An International Symposium*. New York: John Wiley and Sons.
- RAO, J.N.K., et CHOUDHRY, G.H. (1993). Small area estimation: Overview and empirical study. Présenté à l'International Conference on Establishment Surveys, Buffalo, June 1993, à paraître.

où σ_{hkt}^2 est la variance de la population pour l'intersection de la h ème strate et de la k ème petite région. La variance σ_{hkt}^2 peut être évaluée par l'estimateur habituel s_{hkt}^2 pour $n_{hkt} \geq 2$. Notons que l'estimation de la variance conditionnelle v_{kt} fournit aussi une estimation de la variance

inconditionnelle de g_{2kt} .

Si $n_{hkt} = 1$, nous pouvons utiliser une valeur synthétique comme estimation de σ_{hkt}^2 , qu'on peut écrire sous la forme $\sum (n_{hkt} - 1)s_{hkt}^2 / \sum (n_{hkt} - 1)$, la sommation étant faite sur tous les k pour lesquels $n_{hkt} \geq 2$ dans chaque strate (h, t) . Si $n_{hkt} = 0$, v_{ht} dans (A.1) n'est évidemment pas défini. La valeur synthétique de y_{hkt} étant utilisée dans ce cas, il nous faut une valeur synthétique de son erreur quadratique moyenne. Pour chaque (h, t) , celle-ci peut être ainsi définie

$$(\bar{X}_{hkt}/\bar{X}_{ht})^2 (n_{ht}^{-1} - N_{ht}^{-1}) s_{ht}^2 + (\widehat{\text{biais}})^2,$$

où $(\widehat{\text{biais}})^2$ sera défini comme

$$\sum_{n_{ht} > 0} ((\bar{X}_{hkt}/\bar{X}_{ht}) \bar{y}_{ht} - \bar{y}_{ht})^2 / m_{ht},$$

où m_{ht} est le nombre de petites régions avec échantillon dans la strate h .

Selon la notation de (3.2), nous montrons ici l'application de la méthode des moments à l'estimation des composantes de la variance pour le modèle de la section 3.1, dans le cas spécial où il y a une seule variable auxiliaire X_{ht} , $Q_t = \tau^2 I$ et \bar{g}_t suit une marche aléatoire, c.-à-d. $G_{ht}^{(1)} = I$. Soit $F_t = (F_{1t}, \dots, F_{Kt})'$, $F_{kt} = (1, X_{kt})'$, $\bar{g}_t = (\beta_{1t}, \beta_{2t})'$, et $B_t = \text{diag}(\gamma_{1t}^2, \gamma_{2t}^2)$. Le paramètre τ^2 est estimé par la solution de

$$\sum_K \sum_{k=1}^t (g_{2kt} - F_{kt} \bar{g}_t)^2 / (v_{kt} + \tau^2) = T(K - 2).$$

Si l'n'y a pas de solution positive, nous posons $\tau^2 = 0$. Ici, \bar{g}_t désigne l'estimation selon les moindres carrés pondérés de \bar{g}_t , fondée uniquement sur les données transversales au temps t . Cette façon de procéder est semblable à la méthode utilisée dans Fay et Herriot (1979) pour des données transversales. On peut obtenir une estimation de γ_t^2 en résolvant (pour $t = 1, 2$)

$$\sum_T (\beta_{1,t} - \beta_{1,t-1})^2 / (\gamma_t^2 + d_{(t)}'') = T - 1,$$

où $d_{(t)}''$ est le (t, t) ème élément de $(F_{t-1}' U_{t-1}' F_{t-1})^{-1} + (F_t' U_t^{-1} F_t)^{-1}$.

en pratique d'estimer ces biais; toutefois, la taille possible du biais pourrait être évaluée au moyen d'un échantillonnage simulé fait à partir de diverses populations plausibles.

5. CONCLUSION

Une étude de simulation nous a permis de constater que les méthodes d'estimation pour petites régions reposant à la fois sur des données transversales et des données chronologiques peuvent être plus performantes que celles fondées uniquement sur des données transversales, tant du point de vue du biais que de celui de l'erreur quadratique moyenne. Toutefois, le coût en termes de biais pourrait encore être élevé. Il importe évidemment de se demander s'il est possible en pratique de déterminer si les gains découlant de toute forme de lissage l'emportent sur les coûts, et comment faire cette évaluation.

Les modèles visés par l'étude de simulation ont été choisis à partir de considérations générales. Toutefois, en pratique, des diagnostics appropriés semblables à ceux employés dans l'efferrmann et Barnard (1991) devraient être mis au point pour les données d'enquête avant la recommandation de toute méthode assistée par un modèle. Notons également que les estimateurs pour petites régions pourraient être modifiés et rendus robustes à l'égard d'une formulation erronée du modèle sous-jacent, comme il est suggéré dans l'efferrmann et Burck (1990); voir aussi Mantel, Singh et Bureau (1993). Enfin, il y aurait lieu d'étudier, dans de futurs travaux, la modification et l'extension des méthodes présentées dans cet article pour les adapter au cas plus réaliste d'erreurs d'échantillonnage corrélées.

REMERCIEMENTS

Nous tenons à remercier Jon Rao, Danny Pfeffermann et M.P. Singh, avec qui nous avons eu des discussions fructueuses et qui nous ont fourni des commentaires utiles relatifs à des versions précédentes de cet article. Les commentaires et suggestions d'un arbitre anonyme, ainsi que d'un rédacteur associé, ont également été très appréciés. La recherche du premier auteur a été rendue possible en partie par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada accordée à l'Université Carleton.

ANNEXE

A.1 Estimation de la variance de g_{2kt}

Soit v_{kt} la variance conditionnelle (étant donné n_{hkt}) de g_{2kt} dans (2.2). On peut alors exprimer v_{kt} (lorsque $n_{hkt} > 0$ pour tout h) de la façon suivante

$$v_{kt} = \sum^h N_{hkt}^2 \left(n_{hkt}^{-1} - N_{hkt}^{-1} \right) \sigma_{hkt}^2, \tag{A.1}$$

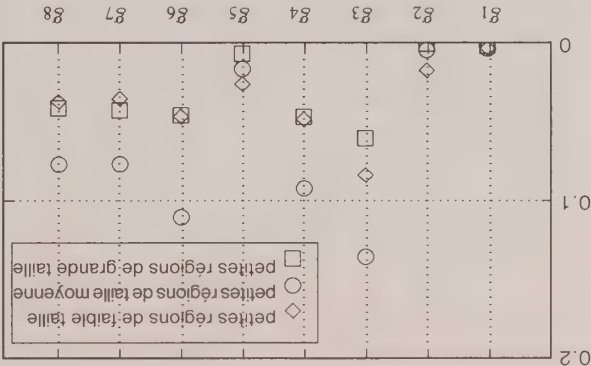
La figure 3 et le tableau 4 sont de format identique à la figure 2 et au tableau 3, mais présentent les biais relatifs plutôt que les racines carrées des erreurs quadratiques moyennes relatives. Tant pour l'estimateur d'expansion g_1 que pour l'estimateur de stratification a posteriori g_2 , les biais sont négligeables. Dans le cas des méthodes de lissage, les biais relatifs absolus pour les petites régions de taille moyenne sont relativement élevés, et sont principalement attribuables aux régions 6 et 8, pour lesquelles la covariable est moins appropriée. Parmi les méthodes de lissage, la méthode liée à la taille de l'échantillon g_5 présente le biais le plus faible, car elle donne généralement des résultats très voisins de ceux de l'estimateur direct g_2 ; toutefois, elle n'est que très peu supérieure à g_2 dans le cas de l'erreur quadratique moyenne. Parmi les autres méthodes de lissage, les estimateurs avec données chronologiques g_7 et g_8 , qui avaient l'erreur quadratique moyenne la plus faible, présentent également le biais le plus faible. Néanmoins, le biais relatif de ces méthodes peut être très élevé, comme dans les régions 6 et 8. Il serait impossible

Biais relatifs absolus et totaux vrais de veaux et bovins pour de petites régions

Tableau 4

Région	Valeurs vraies	Petite taille										Taille moyenne										Grande taille																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																					
		9	10	11	12	Moyenne	1	6	7	8	Moyenne	2	3	4	5	Moyenne	0,042	0,042	0,042	0,042	0,042	0,042																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																					
		8,502	.002	.002	.002	.002	.002	.002	.002	.002	.001	.003	.093	.063	.007	.078	.044	.045	.120	.123	.061	.069	.061	.099	.085	.139	.232	.047	.002	.006	.007	.006	.007	.007	.011	.007	.023	.024	.023	.037	.025	.039	.037	.023	.024	.023	.050	.037	.039																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										
		18,990	.002	.002	.006	.007	.003	.015	.026	.025	.037	.039	.062	.039	.061	.099	.085	.139	.232	.047	.002	.006	.007	.003	.015	.026	.025	.039	.037	.023	.024	.023	.050	.037	.039	.061	.069	.061	.099	.085	.139	.232	.047	.002	.006	.007	.003	.015	.026	.025	.039	.037	.023	.024	.023	.050	.037	.039																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
		19,819	.000	.007	.019	.011	.007	.023	.024	.023	.037	.025	.039	.062	.039	.061	.099	.085	.139	.232	.047	.002	.006	.007	.003	.015	.026	.025	.039	.037	.023	.024	.023	.050	.037	.039	.061	.069	.061	.099	.085	.139	.232	.047	.002	.006	.007	.003	.015	.026	.025	.039	.037	.023	.024	.023	.050	.037	.039																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																
		16,522	.001	.016	.087	.052	.029	.050	.037	.039																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	</

Figure 3. Biais relatifs absolus: moyennes pour des groupes de tailles



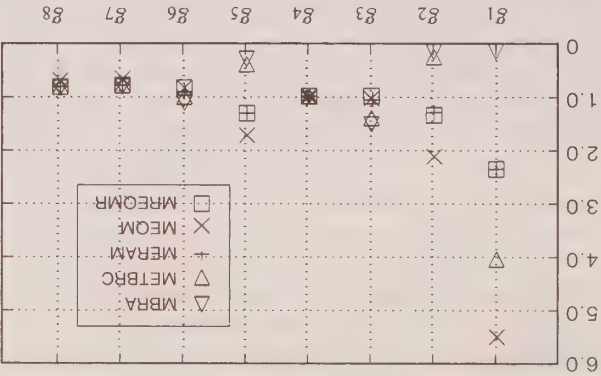


Figure 1. Mesures d'évaluation par rapport à l'estimateur de Fay-Herriot
Nota: METBRC relatif pour g_1 ($= 18,98$) non indiqué.

Moyennes des mesures d'évaluation									
MBRA	METBRC	MERAM	MREQMR	MEQM (000)					
.053	.070	.053	.001	.007	.097	.065	.018	.053	.053
81	82	83	84	85	86	87	88		
.010	.010	.087	.111	.120	.109	.176	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137
.010	.010	.097	.088	.136	.108	.108	.137	.137	.137

se situe généralement entre celle de l'estimateur de Fay-Herriot et celles des estimateurs à données chronologiques g_7 et g_8 . Pour tous les estimateurs (y compris l'estimateur synthétique g_3), l'écart-type du biais relatif conditionnel (METBRC) est appréciable; toutefois, il est le moins élevé dans le cas des méthodes avec données chronologiques. Comme prévu, l'estimateur d'expansion g_1 donne de bons résultats dans le cas de la mesure du biais inconditionnel (MBRA), mais est très peu performant dans le cas de la mesure du biais conditionnel (METBRC). La figure 2 présente les moyennes des mesures $REQM_k$ pour trois groupes de taille, soit les petites régions de faible, moyenne et grande taille, d'après le classement de leurs totaux réels pour la population au temps T . Ces trois regroupements ont été faits parce qu'on s'attendait à ce que les erreurs relatives d'estimation soient supérieures pour les totaux les plus petits, ce que le graphique ne contredit pas. Encore une fois, les méthodes avec données chronologiques g_7 et g_8 donnent les meilleurs résultats. Notons que la méthode avec données chronologiques g_6 , qui suppose que les effets de petite région sont indépendants dans le temps, n'est pas aussi performante. Le tableau 3 présente les mesures $REQM_k$ sans calcul de moyennes. La valeur $REQM_k$ est relativement élevée du fait que le nombre total de vœux et bovins dans la région 9 est inférieur à la moitié de celui de toute autre petite région. Parmi les petites régions de taille moyenne, les régions 6 et 9 se

Figure 2. Racines carrées des erreurs quadratiques moyennes relatives: moyennes pour des groupes de tailles

Région		Valeurs		81		82		83		84		85		86		87		88	
Petite		Moyenne	16,522	.422	.208	.154	.161	.194	.129	.116	.120	.409	.237	.076	.152	.212	.123	.117	.117
Taille moyenne		Moyenne	25,959	.336	.205	.159	.151	.185	.147	.126	.127	.383	.250	.155	.165	.219	.155	.146	.144
Grande		Moyenne	38,642	.259	.164	.098	.100	.149	.083	.083	.084	.270	.176	.091	.097	.160	.088	.085	.088
Petite		9	.580	.277	.342	.275	.277	.199	.160	.174	.104	.360	.196	.078	.113	.175	.097	.103	.087
Petite		10	.18,990	.360	.196	.078	.113	.175	.097	.103	.087	.339	.122	.122	.103	.112	.096	.086	.087
Petite		11	.18,776	.339	.122	.122	.103	.112	.096	.086	.087	.409	.237	.076	.152	.212	.123	.117	.117
Petite		12	.19,819	.409	.237	.076	.152	.212	.123	.117	.117	.336	.205	.159	.151	.185	.147	.126	.127
Taille moyenne		1	.27,595	.312	.206	.117	.130	.185	.120	.100	.102	.306	.241	.256	.216	.224	.168	.172	.172
Taille moyenne		6	.29,012	.306	.241	.256	.216	.224	.168	.172	.172	.341	.121	.107	.094	.110	.088	.092	.092
Taille moyenne		7	.23,600	.341	.121	.107	.094	.110	.088	.092	.092	.383	.250	.155	.165	.219	.155	.146	.144
Taille moyenne		8	.23,627	.383	.250	.155	.165	.219	.155	.146	.144	.422	.208	.154	.161	.194	.129	.116	.120
Grande		2	.35,592	.268	.171	.113	.110	.156	.096	.089	.088	.241	.151	.087	.090	.137	.070	.072	.073
Grande		3	.40,582	.241	.151	.087	.090	.137	.070	.072	.073	.256	.160	.099	.103	.144	.080	.088	.089
Grande		4	.42,396	.256	.160	.099	.103	.144	.080	.088	.089	.270	.176	.091	.097	.160	.088	.085	.088
Grande		5	.35,996	.270	.176	.091	.097	.160	.088	.085	.088	.336	.205	.159	.151	.185	.147	.126	.127
Grande		Moyenne	.38,642	.259	.164	.098	.100	.149	.083	.083	.084	.422	.208	.154	.161	.194	.129	.116	.120

distinguent en étant celles où l'estimation par les méthodes de lissage est la plus difficile. Cela vient du fait que, par rapport à la baisse globale d'environ 16% du nombre total de vœux et bovins dans la pseudo-population entre juin 1986 et janvier 1991, les régions 6 et 8 sont celles qui se situent le plus loin de la moyenne, à 33% et à 1% respectivement, de sorte que la covariable de rajustement par quotient est moins appropriée pour ces régions. Néanmoins, les méthodes avec données chronologiques g_7 et g_8 ont une performance sensiblement meilleure que celle de l'estimateur de stratification a posteriori pour les régions 6 et 8. Cela s'explique par le fait que le modèle de la marche aléatoire, appliqué aux effets de petite région, peut rendre compte de petites régions qui, comme les régions 6 et 8, devient progressivement du modèle.

même configuration d'échantillon (ensemble des tailles des échantillons aléatoires des domaines), que nous appellerons C . La valeur espérée de A étant donné C est une fonction de C , disons $E(C)$, et la variance conditionnelle de A est proportionnelle à m_2^{-1} , disons $V(C)/m_2$. La variance inconditionnelle de A est alors $V\{E(C)\} + E\{V(C)\}/m_2$. m_2 fois est $V\{E(C)\}/m_1 + E\{V(C)\}/m_1m_2$, dont on trouve le minimum, du fait que $m = m_1m_2$ est fixe, en prenant m_1 aussi grand que possible. Pour la deuxième mesure, le choix approprié de m_1, m_2 est moins direct. Dans l'étude de simulation, nous avons fixé m à 30,000 et les valeurs correspondantes de m_1, m_2 étaient de 15,000 et 2.

4.3 Estimateurs utilisés dans l'étude comparative

L'étude a porté sur neuf estimateurs, c'est-à-dire g_1 à g_8 , ainsi que g_7^* , et tous ont été évalués au temps $T = 10$. Nous avons utilisé un modèle de régression linéaire simple pour la composante synthétique, avec variable auxiliaire ainsi définie

$$X_{k^t} = (\theta_1/\theta_1)\theta_{k_1}, \tag{4.6}$$

où θ_{k_1} , θ_1 désignent respectivement les totaux de la population pour la petite région k et la province au temps $t = 1$, c.-à-d. au moment du recensement de 1986. L'estimateur θ_1 dénote l'estimateur de stratification a posteriori de θ_1 d'après les données de l'enquête sur les fermes au temps t , à l'échelle de la province. Ainsi, X_{k^t} est simplement une variable synthétique de rajustement par quotient. Les variances des composantes de l'erreur dans le modèle de régression ont été supposées constantes pour l'ensemble des régions. Pour les modèles chronologiques, nous avons supposé que la dépendance sériale était conforme à une marche aléatoire. Des hypothèses de ce genre ont été utilisées avec succès dans de nombreuses applications, et nous les avons choisies principalement pour favoriser la simplicité. Nous espérons, toutefois, que les modèles choisis conviennent bien au but recherché et puissent illustrer les variations entre les gains obtenus avec différents types d'estimateurs pour petites régions assistés par un modèle, reposant aussi bien sur des méthodes de lissage transversal que sur des méthodes de lissage chronologique.

Puisque les données du recensement de 1986 ont été incluses dans la série chronologique, l'estimation directe \hat{g}_{21} correspond aux résultats du recensement de 1986, et donc l'erreur de l'enquête $\hat{\epsilon}_1$ correspondrait à $\hat{0}$. En outre, d'après la définition de X_{k^t} , un choix raisonnable pour (β_{11}, β_{21}) serait $(0, 1)$, de sorte que \hat{q}_1 doit être $\hat{0}$. Ainsi, les matrices de covariance B_t et W_t au temps $t = 1$ sont nulles, de sorte que la distribution de \hat{q}_t à $t = 1$ n'exige pas d'estimation. Cette modification dans la distribution initiale de \hat{q}_t est naturelle, compte tenu de l'information additionnelle fournie par le recensement. Par ailleurs, puisque les estimations directes \hat{g}_{21} n'étaient pas disponibles pour $t = 2, 3, 4$, les équations d'estimation des composantes de variance du modèle de l'annexe A.2 ont été modifiées en conséquence.

4.4 Résultats empiriques

Les principales conclusions ont été énoncées à la section 1. Nous présentons ici quelques comparaisons détaillées, accompagnées de certaines explications possibles. Nous n'indiquons pas de résultats distincts pour g_7^* , dont les résultats sont un peu inférieurs, quoique dans l'ensemble semblables, à ceux de g_7 . Ces estimateurs sont présentés au tableau 1. Les figures 1 à 3, ainsi que les tableaux 2 à 4, présentent certains résultats empiriques. Nous n'avons pas indiqué les erreurs-types de Monte Carlo, mais toutes étaient largement négligeables.

Tableau 1
Sommaire des estimateurs

g_1 - Expansion	g_2 - Stratification a posteriori	g_6 - MPLSE-I à données chronologiques, les β évoluent dans le temps, les α sont indépendants dans le temps
g_3 - Synthétique		g_7 - MPLSE-II à données chronologiques, les α évoluent dans le temps, β commun et fixe
g_4 - Fay-Herriot		
g_5 - Lié à la taille de l'échantillon		g_8 - MPLSE-III à données chronologiques, les β et les α évoluent dans le temps

Le tableau 2 présente les moyennes pour l'ensemble des petites régions des cinq mesures d'évaluation, et la figure 1 présente sous forme graphique les mesures d'évaluation moyennes par rapport à la valeur pour l'estimateur de Fay-Herriot (g_4). Certaines tendances ressortent clairement des résultats des diverses mesures selon les différents estimateurs. L'estimateur direct g_2 donne un très bon résultat dans le cas de la mesure du biais (MBRA), mais un résultat plutôt médiocre dans le cas des autres mesures. La méthode de lissage transversal g_3 (synthétique) est très peu performante en ce qui a trait aux mesures du biais. La méthode de Fay-Herriot g_4 se révèle un peu meilleure en moyenne que la méthode de stratification a posteriori en ce qui touche la mesure EQM, mais est bien inférieure dans le cas du biais. La méthode liée à la taille de l'échantillon g_5 donne des résultats très semblables à ceux de g_2 ; elle se révèle un peu inférieure dans le cas des mesures du biais et un peu supérieure dans le cas des autres mesures. Les méthodes avec données chronologiques g_7 et g_8 ont une très bonne performance dans l'ensemble, bien qu'elles soient un peu inférieures à g_2 dans le cas du biais. La performance de l'estimateur à données chronologiques g_6

(ii) Écart-type du biais relatif conditionnel pour la région k :

$$ETBRC_k = \left\{ m_1^{-1} \sum_i (B_{ik}^2 - C_{ik}) / \text{vraie}_k - BRA_k^2 \right\}^{1/2}; \quad (4.2)$$

$$B_k = m_2^{-1} \sum_j \text{est}_{jk} - \text{vraie}_k,$$

$$C_{ik} = m_2^{-1} (m_2 - 1)^{-1} \left(\sum_j \text{est}_{jk}^2 - \left(\sum_j \text{est}_{jk} \right)^2 / m_2 \right).$$

Le terme de correction C_{ik} compense le biais de B_{ik}^2 en tant qu'estimation de B_{ik}^2 , du fait que m_2 est fini. $B_{ik}^2 - C_{ik}$ est conditionnellement sans biais pour B_{ik}^2 ; il est aussi inconditionnellement sans biais pour B_k^2 . La moyenne de Monte Carlo $m_1^{-1} \sum_i (B_{ik}^2 - C_{ik})$ converge vers B_k^2 avec probabilité 1 quand $m_1 \rightarrow \infty$. $B_{ik}^2 - C_{ik}$ peut être négatif pour certains i , du fait que m_2 est fini. Si m_1 est grand, la moyenne sur les i est généralement très voisine de B_k^2 ; si la moyenne est inférieure à BRA_k^2 , nous posons $ETBRC_k$ égal à 0. $ETBRC_k$ désignera la moyenne des $ETBRC_k$ pour l'ensemble des régions k .

(iii) L'erreur relative absolue moyenne pour la région k est

$$ERAM_k = m^{-1} \sum_i \sum_j | \text{est}_{ijk} - \text{vraie}_k | / \text{vraie}_k \quad (4.3)$$

et $ERAM$ désigne la moyenne des $ERAM_k$ pour l'ensemble des régions.

(iv) L'erreur quadratique moyenne pour la région k est:

$$EQM_k = m^{-1} \sum_i \sum_j (\text{est}_{ijk} - \text{vraie}_k)^2 \quad (4.4)$$

et EQM , de la même façon, désigne la moyenne pour l'ensemble des régions.

(v) La racine carrée de l'erreur quadratique moyenne relative pour la région k est

$$REQM_k = \{ EQM_k \}^{1/2} / \text{vraie}_k \quad (4.5)$$

et, de la même façon, la moyenne pour l'ensemble des régions est dénotée par $REQMR$.

La précision (c.-à-d. l'erreur-type de Monte Carlo) de chaque mesure dépend de m_1 , m_2 . Pour toutes les mesures, le choix optimal de m_1 , m_2 , avec comme restriction $m_2 > 1$ est $m_1 = m/2$, $m_2 = 2$, car ce choix réduit au minimum l'erreur-type de Monte Carlo. Pour s'en convaincre, posons A comme la moyenne d'une mesure d'évaluation fondée sur m_2 échantillons ayant tous la

stratification optimales pour les strates à tirage partiel. La répartition optimale de Neyman a été utilisée pour déterminer les tailles des échantillons des strates, de façon à optimiser la précision de l'estimation provinciale de la quantité totale. Pour une taille globale d'échantillon de 207 (taux d'échantillonnage de 2%), il en est résulté des tailles de 51, 62, 48 et 35 pour des strates à tirage partiel comprenant respectivement 5,001, 3,188, 1,850 et 312 fermes, tandis que la strate à tirage complet comprenait 11 unités. Le nombre espéré de fermes prélevées dans chaque petite région variait de 4.6 dans la région 9 à 27.5 dans la région 6, la moyenne étant de 17.3. Le nombre espéré de fermes prélevées comptant des vœux et bovins variait de 3.6 dans la région 9 à 18.8 dans la région 3, et la moyenne pour l'ensemble des petites régions était de 11.7. Un total de 30,000 simulations ont été exécutées. À chaque simulation, les échantillons étaient tirés indépendamment pour chaque point dans le temps, au moyen d'un échantillonnage aléatoire simple stratifié sans remplacement. Les 30,000 simulations ont été réalisées sous forme de 15,000 ensembles de 2 simulations, chaque ensemble correspondant à un vecteur différent de tailles d'échantillon réalisées dans les douze petites régions formant chaque strate. Il fallait procéder de la sorte pour calculer certaines mesures d'évaluation conditionnelle, décrites dans la sous-section qui suit; voir aussi Sørndal et Hidiroglou (1989).

4.2 Mesures d'évaluation

Supposons que nous effectuons m simulations dans lesquelles m_1 ensembles de vecteurs différents de tailles d'échantillon réalisées dans les domaines (h, k) sont répétées m_2 fois. Les mesures suivantes peuvent servir à comparer la performance de divers estimateurs au temps T . Supposons que i varie de 1 à m_1 et que j varie de 1 à m_2 .

(i) Biais relatif absolu pour la région k :

$$BRA_k = | m^{-1} \sum_i \sum_j (\text{est}_{ijk} - \text{vraie}_k) / \text{vraie}_k |. \quad (4.1)$$

La moyenne des BRA_k pour l'ensemble des régions k sera désignée par $M BRA$. Nous prenons le biais relatif absolu puisque nous nous intéressons principalement dans cette étude à une mesure globale comme $M BRA$; toutefois, dans d'autres contextes, les biais réels des petites régions individuelles peuvent aussi être d'un intérêt considérable.

La mesure suivante est justifiée par le désir d'évaluer la performance conditionnelle des estimateurs étant donné les vecteurs des tailles d'échantillon réalisées dans les domaines. Il est d'usage de mesurer la performance avec, comme condition, les tailles d'échantillon fixes des domaines; ici, nous considérons l'écart-type du biais conditionnel, B_{ik} , comme une simple mesure sommaire. Si cet écart-type est faible, la méthode est robuste par rapport aux variations des tailles d'échantillon réalisées. Notons que la valeur espérée de B_{ik} est simplement le biais inconditionnel, estimé par BRA_k . Désignons par B_k^2 la valeur espérée inconditionnelle de B_{ik}^2 . Nous définissons la mesure de Monte Carlo suivante:

de fortes corrélations avec la variable correspondante dans le temps au niveau des fermes. Les détails de l'étude empirique sont présentés dans les paragraphes qui suivent.

4.1 Conception de l'étude de simulation

Il nous faut d'abord construire une pseudo-population à partir des données de l'enquête pour six points dans le temps (juin 1988, janvier 1989, . . . , janvier 1991). Le plan réel comporte deux bases (liste et base aréolaire), avec un échantillonnage stratifié à un seul degré pour la liste et un échantillonnage stratifié à deux degrés pour la base aréolaire; pour plus de détails, voir Julien et Maranda (1990). Nous avons décidé de nous limiter aux données provenant de la liste, parce que cette dernière contient les fermes qui existaient au moment du recensement de 1986 et que la variable auxiliaire choisie pour construire le modèle se fonde sur de l'information du recensement de 1986. En outre, nous avons choisi d'utiliser les données de la province de Québec, parce que l'échantillon aréolaire n'y constitue qu'une faible fraction de l'échantillon total et que la variation estimée des coefficients pour les douze districts agricoles (c.-à-d. les petites régions visées) de cette province représentait un vaste intervalle pour les variables relatives au bétail. Pour éviter d'avoir à tenir compte de la variabilité attribuable à l'évolution de la population sous-jacente, nous n'avons conservé que les fermes qui avaient répondu à chacune des six occasions. Par ailleurs, les unités agricoles appartenant à des exploitations multiples à l'une ou l'autre des sept occasions (y compris le recensement) ont été exclues à cause de la difficulté d'extraire les données de chaque ferme à partir des données sommaires des exploitations multiples, et des changements apportés à leurs modalités de déclaration au fil du temps. Les diverses exclusions ci-dessus étaient justifiées par la volonté d'établir des comparaisons plus nettes entre les estimateurs pour petites régions. Le nombre total d'unités agricoles après les exclusions était de 1,160, sur un total de plus de 40,000 fermes figurant dans la liste. Pour la pseudo-population, nous avons reproduit les 1,160 unités agricoles proportionnellement à leur poids d'échantillonnage, de sorte que la taille totale N de la pseudo-population était de 10,362 unités, une quantité se prêtant à une simulation sur micro-ordinateur.

La pseudo-population a été stratifiée en quatre strates à tirage partiel et une strate à tirage complet, la variable de stratification étant les totaux de vœux et bovins du recensement de 1986. Bien que nous n'ayons pas examiné d'autres stratifications ou tailles d'échantillon dans notre étude de simulation, rien ne porte à croire que nos conclusions seraient sensiblement modifiées si nous le faisions. La règle de l'écart sigma (Julien et Maranda 1990) a servi à définir la strate à tirage complet. Pour appliquer la règle de l'écart sigma, nous cherchons la plus petite valeur de la population qui soit supérieure à la médiane et pour laquelle la distance jusqu'à la valeur de la population suivante, en ordre de taille, est d'au moins un écart-type de la population; toutes les unités situées au-delà de ce point sont placées dans la strate à tirage complet. L'algorithme de Sethi (1963) a servi à déterminer les limites de

Il est intéressant de noter que bon nombre des estimateurs examinés jusqu'ici sont optimaux en vertu de cas spéciaux du modèle qui sous-tend \tilde{g}_{87} . Comme nous l'avons vu, on obtient les MPLSE à données chronologiques des méthodes 6 et 7 en imposant des restrictions aux matrices G_i and T_i . Dans le cas des estimateurs de Fay-Herriot transversaux de la section 2.4, les données sont limitées à un seul point dans le temps. Les estimateurs synthétiques de la section 2.3 sont des cas spéciaux des estimateurs de Fay-Herriot avec variance zéro pour les effets aléatoires de petite région, et l'estimateur direct (de stratification a posteriori) est obtenu à la limite, quand la variance des effets de petite région tend vers l'infini. Une autre généralisation pourrait être utile: permettre des corrélations entre les effets de petites régions voisines. Il suffit pour cela que la matrice \tilde{Q}_i de (3.2) puisse être non diagonale; toutefois, il n'est pas évident de savoir quelle serait la structure de corrélation appropriée dans \tilde{Q}_i .

4. ETUDE DE MONTE CARLO

Les méthodes transversales et les méthodes avec données chronologiques ont été comparées de façon empirique au moyen d'une simulation de Monte Carlo. Nous avons utilisé à cette fin une série chronologique réelle tirée des enquêtes bisannuelles sur les fermes de Statistique Canada, soit l'Enquête nationale sur les fermes (en juin) et l'Enquête sur les fermes de janvier. En raison de la refonte qui a suivi le recensement de l'agriculture de 1986, nous avons utilisé les données des six points dans le temps qui étaient disponibles à partir de l'été 1988 pour créer une pseudo-population à des fins de simulation. Nous avons ajouté à cet ensemble des données de l'année du recensement (1986). Nous disposons ainsi de données pour un point additionnel dans le temps, mais il y avait en conséquence trois points manquants dans notre série. Ces vides peuvent toutefois être comblés facilement par des méthodes propres aux séries chronologiques. Notons que la série ainsi obtenue, bien qu'elle soit courte, est jugée adéquate à des fins d'illustration. Le paramètre d'intérêt choisi est le nombre total de vœux et bovins pour chaque district agricole (défini comme la petite région) à chaque point dans le temps. Pour plus de simplicité, des échantillons aléatoires stratifiés indépendants ont été tirés de la pseudo-population pour chaque point dans le temps, bien que les enquêtes sur les fermes utilisent des panels avec renouvellement. Pour modéliser la dépendance des estimations directes des petites régions dans le temps, on a supposé que les totaux globaux sous-jacents des petites régions étaient reliés en vertu d'un quelconque processus aléatoire. La variable auxiliaire utilisée dans le modèle était la valeur rajustée par quotient du total de vœux et bovins du recensement de 1986 dans chaque petite région. Cette variable présentait

Pour trouver l'estimateur optimal (MPLS) de $\tilde{\theta}_T$ dans (3.2) à partir de toutes les estimations directes jusqu'au temps T , nous avons d'abord trouvé le MPLS de $\tilde{\theta}_T$ sous la forme $H_T \tilde{\alpha}_T$. Il est possible, bien que peu pratique, d'obtenir $\tilde{\alpha}_T$ directement à partir des données complètes en recourant à la théorie des modèles linéaires avec effets aléatoires. Toutefois, comme les $\tilde{\alpha}_T$ sont reliés dans le temps selon l'équation de transition (3.2b), il est plus commode de le calculer récursivement au moyen du filtre de Kalman. On a généralement considéré le filtre de Kalman comme une technique bayésienne dans laquelle à chaque temps t , on met à jour la distribution postérieure de $\tilde{\alpha}_t$ pour les données jusqu'au temps $t - 1$ afin d'obtenir la distribution postérieure de $\tilde{\alpha}_t$ pour les données jusqu'au temps t . Bien qu'il soit instructif de percevoir le filtre de Kalman de cette façon, ce n'est pas nécessaire dans le cas de modèles linéaires mixtes. Supposons que $\tilde{\alpha}_{T|s}$ dénote le MPLS de $\tilde{\alpha}_T$ d'après les données jusqu'au temps s , $s < T$. C'est une fait connu (voir Duncan et Horn 1972) que pour la structure spéciale de dépendance sériale considérée ici, le MPLS $\tilde{\alpha}_T$ de $\tilde{\alpha}_T$ d'après les données jusqu'au temps T est le même que le MPLS de $\tilde{\alpha}_T$ d'après $\tilde{\alpha}_{T|s}$ et les $T - s$ dernières observations. En d'autres termes, l'information contenue dans les données précédentes peut être condensée en un MPLS approprié avant que soient ajoutées des données de points dans le temps plus récents. On trouve dans le chapitre 3 de Harvey (1989) une bonne description du filtre de Kalman.

3.1 Méthode 6 (MPLSE-I à données chronologiques)

Pour le premier estimateur, nous laissons $\tilde{\theta}_t$ évoluer dans le temps (selon une marche aléatoire), mais nous supposons que $\tilde{\alpha}_t$ est sériellement indépendant. Les équations du modèle d'états sont dans ce cas semblables à celles de (3.2), sauf que l'indépendance sériale des $\tilde{\alpha}_t$ fait en sorte que $G_{(2)}' = 0$. Il en résulte un estimateur composite

$$\tilde{g}_{6T} = F_T \tilde{\theta}_T + \tilde{\alpha}_T. \quad (3.3)$$

Dans l'étude de simulation décrite plus loin, nous prenons $G_{(1)}' = I$, $B_t' = \text{diag}(\gamma_1^2, \gamma_2^2)$, ce qui correspond à un modèle de marche aléatoire, et $\tilde{Q}_t' = \tau^2 I$. L'annexe A.2 montre comment les paramètres γ_1^2 , γ_2^2 , et τ^2 sont estimés par la méthode des moments. Le filtre de Kalman peut ensuite être appliqué, avec les valeurs initiales de $\tilde{\alpha}_1$ et de son EQM obtenues à partir de l'estimateur FH au temps $t = 1$, ce qui donne le MPLSE de $\tilde{\alpha}_T$. Alors, $H_T \tilde{\alpha}_T$ est l'estimateur MPLSE-I à données chronologiques \tilde{g}_{6T} au temps T .

Comme l'a signalé un arbitre, quand le nombre de petites régions est très élevé, ou quand la variation de $\tilde{\theta}_t$ par rapport à t est relativement grande, il y a peu de différence entre les performances de ces deux estimateurs dans notre étude de simulation décrite à la section 4.

3.2 Méthode 7 (MPLSE-II à données chronologiques)

Pour le deuxième estimateur, nous supposons que $\tilde{\theta}_t$ est fixe (il peut ou non être commun à différents points dans le temps) et que les effets des régions $\tilde{\alpha}_t$ présentent une dépendance sériale conforme, par exemple, à une marche aléatoire. Cette généralisation à des données chronologiques est comparable au modèle proposé par Rao et Yu (1992). L'estimateur composite résultant aura la même forme qu'en (3.1), c.-à-d.

$$\tilde{g}_{7T} = F_T \tilde{\theta}_T + \tilde{\alpha}_T. \quad (3.4)$$

mais les estimations des composantes $\tilde{\theta}_T$ et $\tilde{\alpha}_T$ seront différentes. Deux cas doivent être examinés.

3.2.1 Cas 1: Supposons que les $\tilde{\theta}_t$ sont fixes et invariables dans le temps, mais que les $\tilde{\alpha}_t$ présentent une dépendance sériale. Alors, dans (3.2), $G_{(1)}' = I$ et $B_t' = 0$. Si l'on pose \tilde{Q}_t égal à $\tau^2 I$, le seul paramètre inconnu τ^2 peut être estimé par la méthode des moments; voir l'annexe A.2. Nous désignerons par \tilde{g}_{7T} le MPLSE qui est alors obtenu après substitution de l'estimation du paramètre.

3.2.2 Cas 2: Nous supposons ici que les $\tilde{\theta}_t$ sont fixes, mais qu'ils ne sont pas les mêmes à différents points dans le temps. Les effets des régions $\tilde{\alpha}_t$ évoluent dans le temps comme dans le cas 1. Dans (3.2), nous avons $G_{(1)}' = 0$ et $B_t' = mI$, où m est un nombre élevé. Les expressions de $\tilde{\alpha}_T$ et de son EQM obtenues par le filtre de Kalman dans ce cas donnent les formules correctes quand $m \rightarrow \infty$ (voir Sallias et Harville 1981). Les équations de mise à jour du filtre de Kalman pour $\tilde{\alpha}_t$ ont alors cette forme spéciale

$$\tilde{\theta}_t = (F_t' A_t'^{-1} F_t')^{-1} F_t' A_t'^{-1} (\tilde{g}_{2t} - G_{(2)}' \tilde{\alpha}_{t-1});$$

$$\tilde{\alpha}_t = G_{(2)}' \tilde{\alpha}_{t-1} + P_{t|t-1} A_t'^{-1} (\tilde{g}_{2t} - G_{(2)}' \tilde{\alpha}_{t-1} - F_t' \tilde{\theta}_t);$$

$$P_t = P_{t|t-1} - P_{t|t-1} A_t'^{-1} (A_t' - F_t' A_t'^{-1} F_t')^{-1} F_t' A_t'^{-1} P_{t|t-1}$$

$$A_t'^{-1} P_{t|t-1}$$

où $A_t' = P_{t|t-1} + V_t$, P_t est l'EQM de $\tilde{\alpha}_t$ autour de $\tilde{\alpha}_t$, et $P_{t|t-1} = G_{(2)}' P_{t-1}' + \tilde{Q}_t'$ est l'EQM de $G_{(2)}'$ en tant qu'estimateur de $\tilde{\alpha}_t$. Le MPLSE à données chronologiques, dans ce cas, est désigné par \tilde{g}_{7T} .

3.3 Méthode 8 (MPLSE-III à données chronologiques)

Dans le troisième estimateur, $\tilde{\theta}_t$ et $\tilde{\alpha}_t$ peuvent tous deux évoluer dans le temps. Il en résulte une dépendance sériale plus complexe que dans (3.3) ou (3.4). Cet estimateur aura une forme semblable à celle de (3.1) et peut s'écrire ainsi

$$\tilde{g}_{8T} = F_T \tilde{\theta}_T + \tilde{\alpha}_T. \quad (3.5)$$

2.5 Méthode 5 (estimateur lié à la taille de l'échantillon)

Un autre estimateur composite possible est l'estimateur lié à la taille de l'échantillon de Drew, Singh et Choudhry (1982). Il est défini ainsi

$$\tilde{g}_5 = \Delta \tilde{g}_2 + (I - \Delta) \tilde{g}_3,$$

où $\Delta = \text{diag}(\delta_1, \dots, \delta_K)$,

$$\delta_k = \begin{cases} 1 & \text{si } \sum_{j \in s_k} d_j \geq \lambda N_k, \\ \sum_{j \in s_k} d_j / \lambda N_k & \text{sinon} \end{cases} \quad (2.7)$$

et le paramètre λ est choisi subjectivement comme moyen de contrôler la contribution de la composante synthétique.

L'estimateur ci-dessus tient compte des tailles d'échantillon réalisées n_{hk} , et si ces dernières sont jugées suffisamment élevées selon la condition énoncée en (2.7), il ne dépend pas de l'estimateur synthétique. Cette propriété s'apparente à celle de \tilde{g}_4 ; toutefois, contrairement à \tilde{g}_4 , l'estimateur ci-dessus ne tient pas compte des tailles relatives de la variation intra-région et de la variation inter-régions. Rao et Choudhry (1993) ont démontré de façon empirique comment les MPLSE peuvent parfois se révéler supérieurs aux estimateurs liés à la taille de l'échantillon, notamment quand la variation inter-régions est peu élevée par rapport à la variation intra-région. Sørndal et Hidiroglou (1989) ont également proposé des estimateurs semblables à l'estimateur ci-dessus lié à la taille de l'échantillon.

3. MÉTHODES FONDÉES SUR DES DONNÉES TRANSVERSALES ET CHRONOLOGIQUES COMBINÉES

Supposons qu'on dispose des données recueillies à plusieurs moments différents, $t = 1, \dots, T$, sous forme des estimateurs directs pour petites régions \tilde{g}_{2t} , où \tilde{g}_{2t} est le vecteur des estimations g_{2k} de (2.2) d'après les données dans le temps t , ainsi que des totaux de la population pour la variable auxiliaire dans les petites régions. Nous allons maintenant présenter quelques estimateurs qui généralisent l'estimateur de Fay-Herriot \tilde{g}_{4T} de différentes façons, en tenant compte de la dépendance sériale des estimations directes $\{\tilde{g}_{2t}: t = 1, \dots, T\}$. Rappelons que dans le cas de l'estimateur de Fay-Herriot, le modèle formulé pour $\tilde{\Theta}_T$ a deux composantes: la composante structurelle $F_T \tilde{\beta}_T$ et la composante des régions \tilde{q}_T . L'estimateur \tilde{g}_{4T} emprunte de l'information aux régions pour le moment courant T et est donné par la somme de deux composantes, chacune étant le MPLSE (MELS) pour l'effet aléatoire (fixe) correspondant, c.-à-d.

$$\tilde{g}_{4T} = F_T \tilde{\beta}_T^* + \tilde{q}_T^*. \quad (3.1)$$

Ce modèle appartient à la classe générale définie par Pfeiffermann et Burck (1991) au moyen de modèles chronologiques structurels. Le but principal de leur étude était de démontrer comment la prise en compte des corrélations transversales entre petites régions voisines (en sus des corrélations sériales) et l'inclusion de certaines modifications visant à assurer la robustesse (pour se prémunir contre les défaillances du modèle) pouvaient améliorer la performance des estimateurs fondés sur des modèles chronologiques. Ils ont par ailleurs utilisé la méthode du maximum de vraisemblance dans des conditions de normalité pour estimer les paramètres du modèle. L'objet principal du présent article, en revanche, est l'évaluation de Monte Carlo d'une classe spéciale d'estimateurs à données chronologiques (reliés à l'estimateur de Fay-Herriot) choisis d'après des considérations heuristiques plutôt que par un ajustement de modèle. Les méthodes examinées pourraient donc être considérées comme des méthodes assistées par un modèle, dont la performance sera évaluée dans un cadre fondé sur le plan (c.-à-d. un échantillonnage répété) au moyen d'une simulation de Monte Carlo. De plus, comme nous le verrons plus loin, pour les types de dépendance sériale visés, les paramètres du modèle peuvent être estimés avec une relative simplicité par la méthode des moments, sans qu'il soit nécessaire de faire aucune hypothèse quant à la distribution (p. ex. distribution normale).

$$G_t = \begin{pmatrix} G_{(1)}' & 0 \\ 0 & G_{(2)}' \end{pmatrix}, \quad \tilde{\xi}_t = \begin{pmatrix} \tilde{\xi}_t \\ \tilde{\eta}_t \end{pmatrix}, \quad (3.2c)$$

où

$$\tilde{\alpha}_t = G_t \tilde{\alpha}_{t-1} + \tilde{\xi}_t, \quad (3.2b)$$

et

$$\tilde{g}_{2t} = \tilde{\Theta}_t + \tilde{\epsilon}_t, \quad \tilde{\Theta}_t = F_t' \tilde{\beta}_t + \tilde{q}_t \equiv H_t' \tilde{\alpha}_t \quad (3.2a)$$

Les méthodes fondées sur des données chronologiques pourraient toutelois emprunter également de l'information recueillie à d'autres moments. Nous présentons trois estimateurs qui sont justifiés par des modèles structurels parti-culiers de la dépendance sériale. Ces estimateurs sont tous trois optimaux en vertu de cas spéciaux différents d'un modèle chronologique structurel pour les estimations directes de petites régions $\{\tilde{g}_{2t}: t = 1, \dots, T\}$ définies par le modèle d'espace d'états suivant. Supposons que α_t désigne $(\tilde{\beta}_t', \tilde{q}_t')$ et que H_t' désigne (F_t', I) . On a

Dans nos simulations, les strates a posteriori sont les intersections des strates du plan avec les petites régions, ce qui donne

$$g_{2k} = \sum_h (N_{hk}/n_{hk}) \sum_{j \in s_{hk}} y_{hj} = \sum_h N_{hk} \bar{y}_{hk}. \quad (2.2)$$

Il se peut que cet estimateur ne soit pas lui non plus suffisamment fiable, vu la possibilité que les n_{hk} aient une valeur espérée faible. Si $n_{hk} = 0$, l'estimateur ci-dessus n'est pas défini. Il est d'usage de remplacer \bar{y}_{hk} par 0 lorsque $n_{hk} = 0$. Dans l'étude empirique présentée dans cet article, nous avons remplacé \bar{y}_{hk} par l'estimation synthétique $(X_{hk}/X_h)\bar{y}_h$, où X est une covariable appropriée, dans les cas où $n_{hk} = 0$.

L'estimateur g_{2k} dans (2.2) est conditionnellement sans biais (étant donné $n_{hk} > 0$) et approximativement sans biais en l'absence de condition. L'annexe A.1 donne les détails de l'estimation de l'erreur quadratique moyenne conditionnelle, v_k , de g_{2k} .

2.3 Méthode 3 (estimateur synthétique)

Il est possible de définir un estimateur plus efficace en supposant un modèle qui permette de renforcer les estimations par l'"emprunt" d'information à d'autres petites régions. On obtient ainsi des estimateurs synthétiques; voir par exemple Gonzalez (1973) et Eriksen (1974). Supposons que différents totaux de petites régions soient reliés par une variable auxiliaire X_k selon un modèle linéaire comme

$$\theta_k = \beta_1 + \beta_2 X_k, \quad k = 1, \dots, K, \quad (2.3a)$$

ou, en notation matricielle,

$$\tilde{\theta} = F\tilde{\beta}, \quad (2.3b)$$

où $F = (F_1, F_2, \dots, F_K)'$, $F_k = (1, X_k)'$. Considérons maintenant un modèle applicable aux estimateurs directs pour petites régions g_{2k} , de la forme suivante

$$\tilde{g}_2 = F\tilde{\beta} + \varepsilon,$$

où $\tilde{g}_2 = (g_{21}, \dots, g_{2K})'$, $\tilde{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_K)'$, ε_k sont des erreurs de l'enquête non corrélées avec moyenne 0 et variance v_k . Notons que les g_{2k} ne sont pas corrélés entre les régions puisqu'ils sont conditionnellement (étant donné n_{hk}) sans biais et que les échantillons de petites régions différentes sont conditionnellement indépendants. Si l'on désigne par $\tilde{\beta}$ l'estimation de $\tilde{\beta}$ selon les moindres carrés pondérés (MCP), nous obtenons l'estimateur synthétique de régression de θ_k en vertu du modèle supposé: l'estimateur ci-dessus pourrait être fortement biaisé à moins que le modèle (2.3) ne soit raisonnablement vérifié.

Ce modèle pourrait ne pas être réaliste, car il ne laisse aucune place à des fluctuations aléatoires ou à un effet aléatoire de petite région (disons (a_k) .

2.4 Méthode 4 (estimateur de Fay-Herriot ou MPLSE)

En adoptant l'approche bayésienne empirique de Fay et Herriot (1979) ou l'approche plus générale du meilleur prédicteur linéaire sans biais (voir par exemple Battese, Harter et Fuller (1988), et Pfeffermann et Barnard (1991)), on peut réduire considérablement le biais de l'estimateur synthétique par l'utilisation d'un estimateur composite. Pour une des premières communications touchant l'estimation composite, voir Schabale (1978). L'estimateur composite est obtenu sous forme d'une combinaison convexe de \tilde{g}_2 et d'un \tilde{g}_3 modifié. On suppose à cette fin que

$$\tilde{\theta} = F\tilde{\beta} + \tilde{a}, \quad (2.4)$$

où les a_k sont des effets aléatoires non corrélés de petite région avec moyenne 0 et variance w_k connus jusqu'à une constante. Dans notre étude empirique décrite plus loin, nous prenons $w_k = w$. Notre modèle pour \tilde{g}_2 est donc

$$\tilde{g}_2 = F\tilde{\beta} + \tilde{a} + \varepsilon. \quad (2.5)$$

Ici, on suppose également que \tilde{a} est non corrélé avec ε . Le MPLS de $\tilde{\theta}$ en vertu du modèle défini en (2.4) et (2.5) est

$$\begin{aligned} \tilde{g}_4 &= \tilde{g}_3^* + \Lambda(\tilde{g}_2 - \tilde{g}_3^*) \\ &= \Lambda\tilde{g}_2 + (I - \Lambda)\tilde{g}_3^*, \end{aligned} \quad (2.6)$$

où

$$\begin{aligned} \Lambda &= (V^{-1} + W^{-1})^{-1}V^{-1} = WU^{-1}, \quad U \equiv V + W, \\ V &= \text{diag}(v_1, \dots, v_K), \quad W = \text{diag}(w_1, \dots, w_K), \end{aligned}$$

et $\tilde{g}_3^* = F\tilde{\beta}^*$, $\tilde{\beta}^*$ est l'estimation MCP de $\tilde{\beta}$ en vertu du modèle (2.5). On suppose que les matrices de covariance V et W sont toutes deux connues pour le calcul du MPLS. L'expression (2.6) découle des résultats généraux relatifs aux modèles linéaires avec effets aléatoires; voir par exemple Rao (1973, p. 267) et Harville (1976). Le MPLS ou le MELS (meilleur estimateur linéaire sans biais) de $F\tilde{\beta}$ est \tilde{g}_3^* , et le MPLS de \tilde{a} est $\Lambda(\tilde{g}_2 - \tilde{g}_3^*)$. Il est intéressant de signaler que la structure du MPLS ne change pas peu importe que $\tilde{\beta}$ soit connu ou non. Toutefois, comme il faut s'y attendre, son EQM change, en raison de l'estimation de $\tilde{\beta}$.

Si V et W sont remplacées par des estimations, l'estimateur \tilde{g}_4 est appelé MPLSE. Notons que le modèle (2.4) est plus réaliste que le modèle (2.3) et que, par conséquent, la performance de \tilde{g}_4 devrait être très bonne. L'estimateur \tilde{g}_4 s'approche de \tilde{g}_2 quand les v_k deviennent petits, c.-à-d. quand les n_{hk} deviennent grands. Toutefois, il demeure biaisé, en général, pour $\tilde{\theta}$ donné, le biais tendant vers 0 à mesure que les v_k deviennent petits.

raisonnable pour la grande région, et dans lequel des effets aléatoires de petite région peuvent expliquer tout écart local par rapport au modèle global. Les paramètres de régression et les effets aléatoires de petite région peuvent évoluer dans le temps selon un modèle d'espace d'états également formulé de façon heuristique. Nous n'avons pas examiné ici le problème de l'estimation de l'erreur quadratique moyenne (EQM) de nos estimateurs. Des EQM correspondant aux modèles sous-jacents pourraient être définies et estimées pour bon nombre des estimateurs, mais notre article concerne principalement la performance des estimateurs dans un contexte d'échantillonnage répété. L'estimation de l'EQM est un problème important et difficile, et la disponibilité d'estimateurs fiables de l'EQM pourrait être un facteur important dans le choix des estimateurs. Le but principal de cet article est de comparer des MPLSE utilisant des données chronologiques avec des estimateurs transversaux (p. ex. avec des estimateurs pour domaines avec stratification a posteriori, synthétique, FH et lié à la taille de l'échantillon). Dans la modélisation chronologique des estimations directes des petites régions, nous supposons que les erreurs de l'enquête ne sont pas corrélées dans le temps. Quand les erreurs de l'enquête sont corrélées dans le temps et peuvent être décrites par un modèle raisonnable (p. ex. ARMM), on peut utiliser la méthode de Pfeffermann (1991) pour obtenir des MPLSE à données chronologiques, au moyen du filtre de Kalman. Rao et Yu (1992) obtiennent des MPLSE pour un modèle auquel le filtre de Kalman ne peut s'appliquer, avec des erreurs de l'enquête présentant une structure de corrélation arbitraire dans le temps, mais non corrélées entre les régions. Ils produisent également des approximations du second ordre, ainsi que l'estimation, de l'erreur quadratique moyenne en vertu de leur modèle. Quand il est difficile de formuler un modèle pour les erreurs de l'enquête corrélées, il est possible, au moyen d'un filtre de Kalman convenablement modifié, d'obtenir de bons estimateurs sous-optimaux (Singh et Mantel 1991).

Dans cet article, nous donnons les résultats d'une étude empirique de l'efficacité de MPLSE à données chronologiques. L'étude utilise des simulations de Monte Carlo fondées sur des données de séries chronologiques réelles tirées des enquêtes bisannuelles sur les fermes réalisées par Statistique Canada. Voici les principales conclusions de l'étude:

- (i) Les MPLSE à données chronologiques peuvent produire des gains d'efficacité raisonnables par rapport aux estimateurs transversaux.
- (ii) Dans la classe des méthodes avec données chronologiques examinées dans cet article, il est avantageux de permettre qu'il existe une dépendance sériale dans les effets aléatoires de petite région.
- (iii) Il est prévisible que toute version lissée de l'estimateur direct pour petites régions sera biaisée, mais les MPLSE à données chronologiques affrangent un biais moindre que les méthodes de lissage transversal.

La section 2 présente une description de diverses méthodes transversales servant à l'estimation pour petites régions. Les MPLSE à données chronologiques sont

2. MÉTHODES FONDÉES SUR DES DONNÉES TRANSVERSALES

Dans cette section, nous décrivons certaines méthodes bien connues d'estimation pour petites régions fondées uniquement sur des données transversales. Ghosh et Rao (1994) font un survol intéressant de divers estimateurs pour petites régions.

Soit $\tilde{\Theta}$ le vecteur des totaux Θ_k , $k = 1, \dots, K$ de la population pour de petites régions. Dans la présente section, qui traite des méthodes fondées sur des données transversales, nous ne tenons pas compte du lien entre $\tilde{\Theta}$ et le temps t , à des fins de simplicité.

2.1 Méthode 1 (estimateur d'expansion pour domaines)

Cet estimateur s'exprime ainsi

$$g_{1k} = \sum_{j \in s_k} d_j y_j,$$

où d_j est le poids de l'enquête pour l'unité j de l'échantillon. Dans le cas de l'échantillonnage aléatoire simple stratifié, dont nous servons dans notre étude de simulation à la section 4, nous avons:

$$g_{1k} = \sum_h (N_h/n_h) \sum_{j \in shk} y_j, \quad (2.1)$$

où y_{hj} est la j -ième observation dans la h -ième strate, s_{hk} désigne l'ensemble des n_{hk} unités de l'échantillon se trouvant dans la k -ième petite région et dans la h -ième strate, et n_h , N_h désignent respectivement les tailles de l'échantillon et de la population pour la h -ième strate. Cet estimateur est souvent peu fiable, car n_{hk} , la taille de l'échantillon aléatoire dans la petite région, peut avoir une valeur espérée faible et présenter une variabilité élevée. Avec comme condition la taille d'échantillon réalisée n_{hk} , g_{1k} est biaisé. Toutefois, en l'absence de condition, cet estimateur est non biaisé pour Θ_k .

2.2 Méthode 2 (estimateur pour domaines avec stratification a posteriori)

Nous appellerons aussi cet estimateur l'estimateur direct pour petites régions. Si la taille de la population N_{1k} est connue pour certaines strates a posteriori représentées par l'indice l , l'efficacité de l'estimateur g_{1k} pourrait être améliorée par une stratification a posteriori. Nous définissons

$$g_{2k} = \sum_l N_{lk} \sum_{j \in slk} d_j y_j / \sum_{j \in slk} d_j = \sum_l d_j = \sum_l N_{lk} \bar{y}_{lk}.$$

MPLSE à données chronologiques pour petites régions évaluées à l'aide de données d'enquête

A.C. SINGH, H.J. MANTTEL et B.W. THOMAS¹

RÉSUMÉ

Lorsqu'on effectue des estimations pour petites régions, il est courant de renforcer les estimations en "empruntant" de l'information à d'autres petites régions, car les estimations directes tirées de l'enquête présentent souvent une variabilité d'échantillonnage élevée. Il existe, pour résoudre ce problème de forte variabilité, un ensemble de méthodes dites d'estimation composite, qui utilisent la combinaison linéaire d'un estimateur direct et d'un estimateur synthétique. La composante synthétique se fonde sur un modèle qui relie les moyennes des petites régions transversalement (entre les régions) et/ou par rapport au temps. Le meilleur prédicteur linéaire sans biais empirique (MPLSE) transversal est un estimateur composite fondé sur un modèle de régression linéaire comportant des effets de petite région. Dans cet article, nous examinons trois modèles qui généralisent le MPLSE transversal et permettent d'utiliser les données de plusieurs points dans le temps. Dans le premier modèle, les paramètres de régression sont aléatoires et présentent une dépendance sériale, mais les effets de petite région sont supposés indépendants dans le temps. Dans le deuxième modèle, les paramètres de régression ne sont pas aléatoires et peuvent prendre des valeurs communes dans le temps, mais il y a une dépendance sériale dans les effets de petite région. Le troisième modèle est plus général en ce sens qu'on suppose une dépendance sériale dans les paramètres de régression et dans les effets de petite région. Les estimateurs résultants, ainsi que certains estimateurs transversaux, sont évalués à l'aide des données bisannuelles de l'Enquête nationale sur les fermes et de l'Enquête sur les fermes de janvier de Statistique Canada.

MOTS CLÉS: Estimation composite; modèles d'états; filtre de Kalman; estimateur de Fay-Herriot.

par exemple Bell et Hillimer (1987), Binder et Dick (1989), Pfeffermann (1991) et Tillier (1992).

Dans cet article, nous examinons certaines généralisations naturelles du meilleur prédicteur linéaire sans biais (MPLS) pour petites régions quand on dispose d'une série chronologique d'estimations directes pour les petites régions. L'estimateur de Fay-Herriot (FH), fondé sur le lissage des estimateurs directs par une modélisation transversale des totaux des petites régions, est un important exemple de MPLS pour petites régions. Les estimateurs résultants sont des estimateurs composites (c.-à-d. des combinaisons convexes d'estimateurs directs et synthétiques), et ils sont appelés MPLS empiriques (MPLSE) lorsque des estimations de certaines composantes de variance sont insérées dans l'expression du MPLS. Les travaux de Fay et Herriot (1979) ont marqué une étape importante dans le domaine de l'estimation pour petites régions, car c'est probablement la première fois que celle-ci a été utilisée à grande échelle par des organismes gouvernementaux à des fins d'analyse des politiques. En nous servant de modèles structurels, nous obtenons des MPLSE à données chronologiques, qui comportent à la fois des données transversales et des données longitudinales. Le choix des modèles sous-jacents a été fait d'après des considérations heuristiques générales plutôt qu'en fonction de tests formels des modèles. Les tests formels de ce genre de modèles à l'aide de données d'enquête sont très difficiles et il n'existe pas beaucoup de telles données. Nous commençons plutôt par un modèle de régression qui est

Un grand nombre d'études ont porté sur la façon de produire des estimations pour petites régions au moyen de données d'enquête transversales, en utilisant comme complètement des données du recensement ou de sources administratives. On trouve dans Platak, Rao, Särndal et Singh (1987) un ensemble d'articles instructifs à ce sujet. Les méthodes d'estimation pour petites régions utilisées dans les programmes statistiques fédéraux des États-Unis font l'objet d'une évaluation par le Federal Committee on Statistical Methodology (1993). L'idée de base qui sous-tend toutes les méthodes d'estimation pour petites régions consiste à renforcer les estimations en empruntant de l'information à d'autres petites régions, c'est-à-dire en supposant que différentes régions sont reliées conformément à un modèle contenant des variables auxiliaires tirées des données complémentaires. Par ailleurs, comme il existe de nombreuses enquêtes à passages répétés, il serait également important d'emprunter de l'information à d'autres passages. Des méthodes utilisant des séries chronologiques ont récemment été mises au point dans le but d'améliorer les estimateurs pour petites régions; voir Pfeffermann et Burck (1990) et Rao et Yu (1992). Il est intéressant de signaler qu'après les travaux de Scott et Smith (1974) sur le renforcement des données d'enquête par le recours à des séries chronologiques, on n'a constaté que récemment un regain d'intérêt pour l'estimation d'aggrégats tirés d'enquêtes complexes répétées à intervalles réguliers; voir

¹ A.C. Singh et H.J. Mantel, Division des méthodes d'enquêtes sociales; B.W. Thomas, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6.

cette affirmation mais en comparant les données des tableaux 2 et 5, on peut avoir une idée de la taille du biais qui serait nécessaire pour que l'estimateur direct et l'estimateur fondé sur un modèle aient la même EQM. Prenons l'estimation du nombre total de consommateurs pour le produit de type E; cette estimation est fortement influencée par la procédure et donc, probablement très susceptible d'être entachée d'un biais. La variance (et donc l'EQM) de l'estimateur direct est $1,146^2 = 1,313,316$, tandis que celle de l'estimateur fondé sur un modèle est $675^2 = 455,625$. Par conséquent, il faudrait que la valeur estimée de 7,366 (selon un modèle) soit entachée d'un biais de 926 pour que l'EQM soit la même pour les deux estimateurs.

REMERCIEMENTS

Les auteurs tiennent à remercier les arbitres pour leurs commentaires utiles.

BIBLIOGRAPHIE

DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. *Techniques d'enquête*, 8, 19-52.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

GHOSH, M., et RAO, J.N.K. (1994). Small area estimation: An appraisal. À paraître dans *Statistical Science*.

GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.

GONZALEZ, M.E., et HOZA, C. (1978). Small area estimation with application to Unemployment and Housing Estimates. *Journal of the American Statistical Association*, 73, 7-15.

HOLT, D., et HOLMES, D.J. (1993). Small domain estimation for unequal probability survey designs. Working Paper Series, No. 2, Department of Social Statistics, University of Southampton, UK.

HOLT, D., et SMITH, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society, Sér. A*, 142, 33-46.

PLATEK, R., RAO, J.N.K., SÄRNDAAL, C.-E., et SINGH, M.P. (1987). *Small Area Statistics*. New York: John Wiley and Sons.

SÄRNDAAL, C.-E., et HIDRIGLOU, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAIKLE, W.L., BROCK, D.B., et SCHNACK, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the Social Statistics Section, American Statistical Association*, 1017-1021.

6. ANALYSE

Si nous comparons ces résultats aux estimations du tableau 2, nous observons une forte réduction des erreurs types des totaux estimés (de 40 à 80%) sauf pour le produit de catégorie B. De la même manière, les erreurs types des parts de marché estimées sont beaucoup moins élevées avec un modèle.

Ces estimations sont accompagnées des erreurs types en p correspondantes.

Le tableau 5 contient aussi les estimations (basées sur un modèle) du nombre total d'automobiles de type k achetées, $\theta_k(3)$, et de la part de marché correspondante, $\hat{\theta}_k(3)$, calculées à l'aide des équations (17) et (19) respectivement, pour la classification définie dans le tableau 4.

5.2 Estimation du nombre d'automobiles

En ce qui concerne l'estimation du taux de pénétration du marché, on observe aussi dans ce cas-là des erreurs types moins élevées pour tous les produits (des écarts de 30 à 40% sauf pour les produits A et B, pour lesquels les écarts sont un peu moins prononcés).

Si les hypothèses d'indépendance conditionnelle ne se vérifient pas, les estimateurs comporteront un biais résiduel au sens normal du plan, mais ce peut être un risque acceptable pour garantir la stabilité des estimateurs pour l'ensemble des produits. En ce qui concerne les résultats numériques observés dans les sections précédentes, seules les estimations basées sur un modèle relatives au produit B sont en dehors de l'intervalle de confiance à 95% basé sur l'estimateur direct. Les hypothèses d'indépendance conditionnelle dépendent du choix de la catégorie f et elles peuvent être testées à l'aide de tests chi carré pour tableaux de contingence.

Bien que les données du tableau 5 montrent que les erreurs types de plan des estimations basées sur un modèle sont généralement moins élevées que les erreurs types des estimations directes (tableau 2), on peut prétendre que les estimateurs fondés sur un modèle sont susceptibles d'être biaisés et que par conséquent, leur erreur quadratique moyenne (EQM) n'est pas moins élevée que celle des autres estimateurs. Le biais des estimateurs fondés sur un modèle viendra de l'inadéquation des hypothèses d'indépendance conditionnelle (par ex., équation (5)). On ne peut vérifier

5. RÉSULTATS EMPIRIQUES

5.1 Estimation du nombre de consommateurs

Dans la section 4.2, nous avons étudié l'efficacité de $\hat{\theta}_k(2)$ pour diverses structures de population pour lesquelles l'hypothèse (5) se vérifiait. Cet exercice peut paraître douteux aux yeux du lecteur puisque l'hypothèse (5) ne se vérifie pas dans la pratique. Nous allons maintenant nous servir des données réelles d'enquête pour calculer $\hat{\theta}_k(2)$ pour une classification spéciale basée simultanément sur deux critères: la taille du parc d'automobiles et un indicateur permettant de savoir si le consommateur a acheté ou non des automobiles d'un type quelconque à des fins commerciales (voir tableau 4). Schabile, Brock et Schnack (1977) et Drew, Singh et Choudhry (1982) ont fait des analyses empiriques d'estimateurs synthétiques pour différentes situations.

Pour chacun des types de produit énumérés dans le tableau 2 (A à G), nous avons soumis à un test χ^2 l'hypothèse qu'étant donné la catégorie de la variable de condition (f), la décision d'un consommateur d'acheter le produit ne dépend pas de la strate (h). Notons que pour notre exemple, nous utilisons un plan d'échantillonnage aléatoire stratifié et les hypothèses de distribution multinomiale habituelles s'appliquent. Dans le cas de plans à plusieurs degrés, il faudrait modifier l'analyse chi carré ordinaire en utilisant par exemple des facteurs de correction de Rao-Scott. En pratique, il est difficile de trouver une classification de f qui soit telle que les hypothèses d'indépendance conditionnelle (5) se vérifient pour chaque type de produit. Néanmoins, pour la classification définie dans le tableau 4, nous avons observé que la variation de la probabilité qu'on achète un produit d'un type donné était expliquée en majeure partie par la catégorie de la variable de condition (f) et qu'une infime partie seulement de la variation des résiduels était attribuable à des différences entre les strates.

Tableau 4		
Définition des catégories (f) de la variable de condition		
Catégories	Taille du parc	Achat de véhicules
Définition de f		
1	Néant	0
2	1-4	> 0
3	5-8	> 0
4	9-15	> 0
5	16-25	> 0
6	26-50	> 0
7	51-100	> 0
8	101-200	> 0
9	201-550	> 0
10	> 550	> 0

Les estimations basées sur un modèle calculées à l'aide des équations (6) et (9) – c'est-à-dire les estimations relatives aux consommateurs $\hat{\theta}_k(2)$ et $\hat{\eta}_k(2)$ – sont reproduites dans le tableau 5. Les variances basées sur un modèle risquent d'exagérer la précision des estimateurs étant donné qu'elles reposent sur les hypothèses d'indépendance conditionnelle du modèle et que celles-ci peuvent être fausses dans la pratique. En revanche, il est toujours possible de calculer l'estimation empirique de la variance en p de l'estimateur fondé sur un modèle (voir Holt et Holmes 1993). Ce calcul ne nécessite aucune hypothèse de distribution ou d'indépendance conditionnelle et son résultat peut être considéré comme une mesure plus objective. Les estimations de l'erreur type calculées de cette manière figurent dans le tableau 5. Comme il s'agit d'estimations fondées sur un plan, elles renferment des corrections pour population finie. [Notons ici que les erreurs types basées sur un modèle pour $\hat{\theta}_k(2)$ (qui ne figurent pas dans le tableau 5) étaient environ 10% moins élevées que les erreur types fondées sur un plan pour tous les cas.]

Tableau 5

Estimations basées sur un modèle et erreurs types en p correspondantes pour certains types de produits

Produit (k)	Estimation du nombre de consommateurs		Estimation du nombre d'automobiles	
	Total	Pénétration $\hat{\eta}_k(2)$	Total	Part de marché $\hat{\eta}_k(3)$
A	63,433	.4070	263,511	.3722
B	39,673	.2546	177,667	.2501
C	21,930	.1407	65,357	.0923
D	13,422	.0861	22,146	.0313
E	7,366	.0473	15,798	.0223
F	5,826	.0374	14,398	.0203
G	7,686	.0493	11,207	.0158
1 ^{ère} ligne: estimation		(.0039)	(813)	(.0011)
2 ^{ème} ligne: e.r. en p				

Si nous comparons ces résultats à ceux du tableau 2, nous observons que les erreurs types des totaux estimés sont beaucoup moins élevées: de 30 à 40% moins élevées pour tous les produits sauf A et B (les principaux modèles), pour lesquels la différence est plutôt de 15 à 20%. Ces observations étaient prévisibles puisque le plan de sondage de l'enquête est optimal pour le total des ventes d'automobiles et donc, relativement efficace pour des produits qui occupent une grande part de marché. Par conséquent, la méthode fondée sur un modèle devrait être utile surtout pour des produits dont la part de marché est plus modeste.

où $N_{fm}^h = \sum_h N_{h n_{hfm}/n_h}$, et $y_{kfm}^h = \sum_h y_{kfm}^h/n_{hfm}$ est le nombre moyen non pondéré (pour l'échantillon) d'automobiles de type k achetées par des consommateurs de la catégorie f qui ont acheté au total m automobiles de toute sorte. Les probabilités de sélection servent ici à définir un estimateur pondéré de N_{fm}^h , c'est-à-dire le nombre total de consommateurs de la catégorie f qui achètent m voitures de toute sorte. Cet estimateur a une forme semblable à celle de l'estimateur défini en (6). Suivant les hypothèses de modèle (13), on peut montrer que

$$V_k(\hat{\theta}_k(2)) = \sum_h \sum_f \sum_m \frac{N_h^2}{N_{fm}^h} \mu_{fm}^h \hat{\theta}_{fm|h} (1 - \hat{\theta}_{fm|h}) \quad (14)$$

$$- \sum_h \sum_f \sum_m \sum_{f' \neq f} \sum_{m' \neq m} \frac{N_h^2}{N_{fm}^h} \mu_{fm}^h \mu_{f'm'}^h \hat{\theta}_{fm|h} \hat{\theta}_{f'm'|h}$$

$$+ \sum_h \sum_f \sum_m \frac{N_h^2}{\sigma_{fm}^2 \hat{\theta}_{fm|h}} \sum_h n_h \hat{\theta}_{fm|h} \left\{ (1 - \hat{\theta}_{fm|h}) + n_h \hat{\theta}_{fm|h} \right\}$$

$$+ \frac{\sum_h n_h \hat{\theta}_{fm|h}}{[1 + (2n_h - 3)\hat{\theta}_{fm|h} - 2(n_h - 1)\hat{\theta}_{fm|h}^2]} \quad (15)$$

où $\hat{\theta}_{fm|h} = \hat{\theta}_{f|h} S_{m|h_f} / \mu_{fm}^h = E_k\{Y_{kfm|h}\}$, et $\sigma_{fm}^2 = V_k\{Y_{kfm|h}\}$.

Dans la pratique, y_{kfm}^h reposera sur un très petit nombre d'observations si seulement quelques consommateurs de la catégorie f achètent exactement m voitures. Pour obtenir une plus grande stabilité, on pourrait définir m comme une variable ordinaire en créant un petit nombre de classes pour le total d'automobiles achetées. De cette manière, l'hypothèse (13) signifierait que la distribution du nombre d'automobiles de type k achetées est la même à l'intérieur d'une catégorie f et d'une classe m . En outre, on pourrait considérer ℓ comme une variable aléatoire continue et faire des hypothèses sur sa distribution, ce qui nous amènerait à des estimateurs par quotient ou à des estimateurs par régression. Un second ensemble d'hypothèses paramétriques, plus contraignantes encore, est le suivant:

$$T_{\ell|hfm} = T_{\ell|fm} \quad \text{pour tous } h, \quad (16)$$

$$S_{m|h_f} = S_{m|f} \quad \text{pour tous } h.$$

Ces hypothèses signifient qu'étant donné la catégorie f , la distribution conjointe du nombre d'automobiles de type k achetées et du nombre total d'automobiles achetées, tous modèles confondus (m) est indépendante de la strate (h). Dans ces conditions, l'estimateur du maximum de vraisemblance de θ_k est défini par l'équation

$$\hat{\theta}_k(3) = \sum_f N_f y_{kf}, \quad (17)$$

où $y_{kf} = \sum_h y_{kf}^h/n_f$ est le nombre moyen non pondéré (pour l'échantillon) d'automobiles de type k achetées par des consommateurs de la catégorie f , peu importe le nombre total d'automobiles achetées, et $N_f = \sum_h N_h n_{hf}/n_h$ est l'estimateur pondéré du nombre total de consommateurs inclus dans la catégorie f . On peut montrer que, suivant les hypothèses (16),

$$V_k(\hat{\theta}_k(3)) = \sum_h \sum_f \frac{N_h^2}{N_{fm}^h} \mu_{fm}^h \hat{\theta}_{fm|h} (1 - \hat{\theta}_{fm|h})$$

$$- \sum_h \sum_f \sum_{f' \neq f} \frac{N_h^2}{N_{fm}^h} \mu_{fm}^h \mu_{f'm'}^h \hat{\theta}_{fm|h} \hat{\theta}_{f'm'|h}$$

$$+ \sum_h \sum_f \frac{N_h^2}{\sigma_{fm}^2 \hat{\theta}_{fm|h}} \sum_h n_h \hat{\theta}_{fm|h} \left\{ (1 - \hat{\theta}_{fm|h}) + n_h \hat{\theta}_{fm|h} \right\}$$

$$+ \left[\frac{\sum_h n_h \hat{\theta}_{fm|h}}{1 + (2n_h - 3)\hat{\theta}_{fm|h} - 2(n_h - 1)\hat{\theta}_{fm|h}^2} \right] \quad (18)$$

Si les hypothèses (16) étaient plausibles, y_{kf}^h reposerait sur un plus gros échantillon que y_{kfm}^h (équ. 14) et $\hat{\theta}_k(3)$ serait donc plus stable.

L'estimateur du maximum de vraisemblance de la part de marché pour le produit de type k , R_k , suivant l'hypothèse (16), est défini

$$\hat{q}_k(3) = \frac{\sum_f N_f y_{kf}}{\sum_f N_f z_{kf}}, \quad (19)$$

où z_{kf} , défini comme y_{kf}^h , est le nombre moyen non pondéré (pour l'échantillon) d'automobiles de toutes sortes achetées par des consommateurs de la catégorie f .

($f = h$) et le reste dans les catégories voisines (pour (b) et (c) par exemple, $f = h - 1, h + 1$). Enfin, la structure (e) suppose que, pour n importe quelle strate h , les consommateurs ont autant de chances d'appartenir à l'une ou l'autre des catégories $f = 1, \dots, 10$.

En ce qui regarde $P_{k|f}$, la structure (i) suppose un modèle de voiture qui est acheté peu fréquemment par les consommateurs qui ont parc d'automobiles réduit (c.-à-d. ceux qui appartiennent à la catégorie 1 ou 2), et qui n'est pas du tout acheté par les consommateurs qui ont de grands parcs d'automobiles. La structure (iii) suppose un modèle de voiture qui est acheté selon une probabilité qui est inversement proportionnelle à la taille du parc d'automobiles, tandis que la structure (iv) suppose un modèle de voiture qui est acheté selon une probabilité directement proportionnelle à la taille du parc d'automobiles. En ce qui concerne la structure (iv), un modèle est acheté dans une probabilité de 0,5, peu importe la taille du parc d'automobiles du consommateur.

Le tableau 3 donne les valeurs du facteur d'efficacité exprimé par l'équation (10) pour chaque combinaison de structures de $Q_{f|h}$ et de $P_{k|f}$ selon la formule de répartition non proportionnelle définie dans le tableau 1. La colonne (a) du tableau correspond au cas dans lequel les catégories f coïncident avec les strates, et les deux estimateurs, $\hat{\theta}_k(1)$ et $\hat{\theta}_k(2)$, sont identiques. Le tableau montre qu'on peut réaliser de forts gains d'efficacité (par ex., 70%) pour certaines combinaisons de paramètres: moins il y a de rapport entre f et h , plus le gain d'efficacité est élevé. Même dans le cas des structures (c) et (d), où f et h sont étroitement liés, on peut observer des gains d'efficacité appréciables. Les gains d'efficacité dépendent beaucoup plus de la structure de $Q_{f|h}$ que de celle de $P_{k|f}$.

Tableau 3

Facteurs d'efficacité, $e[\hat{\theta}_k(2)]$, pour diverses combinaisons de structures de $Q_{f|h}$ et de $P_{k|f}$

Structure de $Q_{f h}$		(a)	(b)	(c)	(d)	(e)
Structure de $P_{k f}$	(i)	0	0.108	0.196	0.355	0.648
	(ii)	0	0.116	0.206	0.391	0.695
	(iii)	0	0.103	0.181	0.387	0.695
	(iv)	0	0.115	0.203	0.391	0.706

Dans le cas de la structure (e), où $Q_{f|h}$ a la même valeur pour tous f et h , il est possible de montrer que le facteur d'efficacité peut être exprimé par l'équation

$$e[\hat{\theta}_k(2)] = \left(1 - \frac{P_{k|f}(1 - P_{k|f})}{\delta^2} \right) \frac{\sum_h \tau_h N_h^2 / n_h}{\sum_h N_h^2 / n_h}, \quad (11)$$

sont, respectivement, la moyenne et la variance de $\{P_{k|f}\}$ pour l'ensemble des catégories f et $\tau_h = 1 - n_h/n + O(n^{-1})$. Le terme entre parenthèses dans l'équation (11) peut prendre des valeurs de 0 à 1, selon la distribution des valeurs de $\{P_{k|f}\}$ pour les catégories f . Dans le cas (iv), $P_{k|f}$ est constant et le terme entre parenthèses est donc égal à un. Le second terme du membre de droite de l'équation (11) dépend uniquement du plan de sondage et sa valeur pour le mode de répartition de l'échantillon défini dans le tableau 1 est 0,706.

4.3 Estimation de X_k : le nombre d'automobiles de type k achetées

On peut se servir de la méthode exposée dans la section 4.1 pour estimer le nombre d'automobiles achetées. Nous ajoutons une variable de condition pour représenter le nombre total d'automobiles achetées, m , tous modèles confondus, et nous donnons à la variable aléatoire une nouvelle extension, X_{khfmi} , c'est-à-dire le nombre d'automobiles de type k achetées par le consommateur i de la strate h et de la catégorie f qui achète m voitures de toutes sortes. Ainsi, le nombre d'automobiles de type k achetées variera en fonction du nombre total d'automobiles achetées. Posons

$$S_{m|h} = \text{Prob}\{\text{consommateur achète } m \text{ automobiles de toute sorte} \mid h, f\}, \quad m = 0, 1, 2, \dots,$$

$$T_{\ell|hfm} = \text{Prob}\{\text{consommateur achète } \ell \text{ automobiles de type } k \mid h, f, m\}, \quad \ell = 0, 1, \dots, m.$$

On peut alors exprimer le paramètre à l'étude basé sur un modèle, qui est l'équivalent du nombre total de produits de type k achetées (X_k), par l'équation

$$\theta_k = \sum_h \sum_f \sum_m \sum_{\ell} N_h Q_{f|h} S_{m|h} T_{\ell|hfm}, \quad (12)$$

qui est une extension de l'équation (4). Nous considérons deux ensembles d'hypothèses additionnels, dont le premier est

$$T_{\ell|hfm} = T_{\ell|f} \quad \text{pour tous } h. \quad (13)$$

Ces hypothèses signifient qu'étant donné la catégorie f et le nombre total d'automobiles neuves achetées, m , la distribution du nombre d'automobiles de type k achetées est indépendante de la strate (h).

L'estimateur du maximum de vraisemblance de θ_k , suivant les hypothèses (13), est

$$\hat{\theta}_k(2) = \sum_f \sum_m N_{f|m} \bar{y}_{kf|m}, \quad (14)$$

Si N_{hf} – le nombre total de consommateurs inclus dans la strate h et la catégorie f (taille du parc d'automobiles) – est connu, on peut élargir l'équation (1) de la manière habituelle et exprimer le paramètre étudié sous cette nouvelle forme:

$$\Theta_k = \sum_h \sum_f N_{hf} P_{k|hf}. \quad (3)$$

L'équation ci-dessus équivaut à la stratification a posteriori lorsque $\{N_{hf}\}$ est connu et si tel est le cas, l'information supplémentaire entraînera un gain d'efficacité (Holt et Smith 1979). Si $\{N_{hf}\}$ est inconnu, nous pouvons reformuler le modèle selon deux ensembles de probabilités:

$$\tilde{Q}_{f|h} = \text{Prob} \{ \text{consommateur possède un parc de taille } f \mid \text{strate } h \},$$

$$P_{k|hf} = \text{Prob} \{ \text{consommateur achète un produit de type } k \mid \text{strate } h \text{ et taille } f \text{ du parc de véhicules} \}.$$

On peut alors exprimer le paramètre étudié par l'équation

$$\Theta_k = \sum_h \sum_f N_h \tilde{Q}_{f|h} P_{k|hf}. \quad (4)$$

Nous pouvons obtenir un autre estimateur fondé sur un modèle en faisant d'autres hypothèses sur les paramètres de modèle. Supposons maintenant que

$$P_{k|hf} = P_{k|f} \quad \text{pour tous } h. \quad (5)$$

Cette équation implique qu'étant donné la catégorie f (la taille du parc d'automobiles qu'exploite un consommateur), la probabilité que le consommateur achète un produit de type k est *indépendante* de la strate (h) à laquelle appartient ce consommateur. Du point de vue algébrique, l'hypothèse est semblable à celle utilisée dans l'estimation synthétique pour petites régions sauf que dans ce dernier cas, on regroupe des données de plusieurs régions, ce qui est inacceptable dans le cas qui nous occupe. Nous allons plutôt regrouper des données de plusieurs strates dans le domaine étudié. Il s'agit en fait de choisir une variable de condition qui tienne compte de la relation marginale entre la décision d'un consommateur d'acheter un produit de type k et la strate à laquelle appartient ce consommateur.

En se servant de l'hypothèse (5) et en élargissant la notation de la façon habituelle ($n_{kf} = \sum_h n_{khf}$, etc.), on peut montrer que

$$\tilde{Q}_{f|h} = \frac{n_h}{n_{hf}}, \quad P_{k|f} = \frac{n_{kf}}{n_f}$$

et l'estimateur du maximum de vraisemblance de Θ_k devient

$$\hat{\Theta}_k(2) = \sum_h \sum_f N_h \frac{n_{hf}}{n_{hf}} \frac{n_h}{n_f} = \sum_f N_f \frac{n_{kf}}{n_f} \quad (6)$$

(6)

où $N_f = \sum_h N_h n_{hf}/n_h$ et $\hat{Y}_{kf} = n_{kf}/n_f$ est la moyenne d'échantillon non pondérée pour les consommateurs de la catégorie f (c.-à-d. la proportion empirique de consommateurs de la catégorie f qui achètent un produit de type k). Par conséquent, l'équation (6) a la forme d'un estimateur de stratification basé sur la classification f , sauf que l'effectif de population des strates, $\{N_f\}$, est inconnu. Notons qu'on aurait eu naturellement un estimateur du même type, à la différence que $\{N_f\}$ serait connu, si un échantillon stratifié basé sur f avait été prélevé. En fait, ce n'est pas le cas: les membres de la catégorie f qui sont échantillonnés ne le sont pas avec la même probabilité. Cependant, à cause des hypothèses paramétriques, on considère l'échantillon dans chaque catégorie f comme un échantillon à probabilités égales puisque suivant l'hypothèse (5), les poids de sondage sont non informatifs et amènent simplement une perte d'efficacité lorsqu'on estime $P_{k|f}$. Par conséquent, même si les fractions de sondage n_h/N_h servent à estimer $\{N_f\}$, elles ne sont pas utilisées explicitement dans $\hat{P}_{k|f} = n_{kf}/n_f = \hat{Y}_{kf}$. Notons que l'estimateur regroupe de l'information provenant des strates h à l'intérieur du domaine k mais non dans plusieurs domaines à la fois (c.-à-d. les produits).

Il convient de noter que si n_h/N_h est constant, l'équation (6) se ramène à l'estimateur par facteur d'extension habituel défini en (2) et l'hypothèse (5) ne donne pas de nouvel estimateur. Si l'échantillon est réparti de façon non proportionnelle, l'hypothèse amène l'utilisation des poids de sondage pour N_f (où ils sont nécessaires) mais non pour l'estimation de $P_{k|f}$ (où ils sont non informatifs étant donné f et l'hypothèse (5)).

L'équation (5) représente un ensemble d'hypothèses contraignantes, qui exigent que les valeurs $P_{k|hf}$ soient égales à une valeur unique $P_{k|f}$ pour tous h . En pratique, on peut introduire des hypothèses aléatoires telles que $P_{k|hf} = P_{k|f} + \epsilon_{k|hf}$, où $E[\epsilon_{k|hf}] = 0$ et $V[\epsilon_{k|hf}] = \sigma_{\epsilon}^2$.

Ces hypothèses conduiront à une analyse hiérarchique de Bayes ou à une analyse empirique de Bayes, comme le décrivent Ghosh et Rao (1994) ou Fay et Herriot (1979). Nous n'examinons pas ici ces méthodes car cela évaquerait la forme élémentaire de l'estimateur fondé sur un modèle de même que l'éclairage qu'elle nous apporte. Dans le même esprit, on peut appliquer la méthode de Särndal et Hidiroglou (1989) ou celle de Drew, Singh et Choudhry (1982) pour obtenir des estimateurs dépendants de la taille d'échantillon sans contrevenir à la règle voulant que l'estimateur regroupe de l'information qui provient d'un seul domaine et non de plusieurs domaines à la fois (produits).

Nous pouvons comparer les estimateurs (2) et (6) lorsque l'hypothèse (5) se vérifie puisqu'on peut montrer que

pour la stratification a posteriori, l'estimation par quotient ou l'estimation par régression mais dans tous ces cas, il faut connaître les moyennes ou les totaux de population. Or, cette information n'existe pas. C'est pourquoi nous recou- rons plutôt à une méthode fondée sur un modèle pour obtenir d'autres estimateurs pour toute la gamme de produits.

4.1 Estimation de Y_k^* : le nombre de consommateurs qui achètent un produit de type k

Nous considérons en premier lieu le nombre de consom- mateurs qui achètent un produit de type k . Nous donnons à Y_k^* une nouvelle extension, Y_{kh}^* , dans le but évident de définir la variable aléatoire indicatrice d'achat pour le produit k et le consommateur i dans la strate h . Nous considérons chaque décision de consommateur comme le résultat d'un tirage bernoullien. Soit P_{kh} la probabilité qu'un consom- mateur dans la strate h achète une automobile de type k [$P_{kh} = \text{Prob}(Y_{khi}^* = 1)$]. Nous définissons l'équivalent de Y_k^* pour un modèle, c.-à-d. le nombre total de consom- mateurs du produit de type k , par la formule

$$\Theta_k(1) = \sum_h N_h P_{kh} \tag{1}$$

En supposant que chaque décision de consommateur est indépendante, on peut exprimer la vraisemblance comme le produit habituel de termes binomiaux. Les esti- mateurs du maximum de vraisemblance sont définis par l'équation $P_{kh} = n_{kh}/n_h$ et l'estimateur du maximum de vraisemblance de Θ_k est l'estimateur de plan de sondage stratifié bien connu

$$\Theta_k(1) = \sum_h \frac{n_h}{N_h} n_{kh} = \sum_h N_h \hat{y}_{kh} \tag{2}$$

où n_{kh} est le nombre de consommateurs échantillonnés de la strate h qui achètent un produit de type k , n_h est l'effectif de la strate pour l'échantillon et $\hat{y}_{kh} = n_{kh}/n_h$ est la moyenne d'échantillon pour les consommateurs de la strate h (c.-à-d. la proportion empirique de consom- mateurs de la strate h qui achètent un produit de type k). L'estimateur ci-dessus est peu satisfaisant en règle générale lorsqu'un trop petit nombre de consommateurs achètent le produit de type k .

Supposons que nous introduisions une autre variable de condition, par laquelle chaque consommateur peut être classé dans une catégorie f , $f = 1, \dots, F$, et que nous donnions à la variable aléatoire indicatrice une nouvelle extension, Y_{khi}^* . Les catégories f recouperont les strates h et il s'agit de définir f de telle sorte que, pour n'importe quelle catégorie donnée, la décision d'un consommateur d'acheter le produit de type k ne dépende pas de la strate à laquelle appartient ce consommateur. En ce qui concerne les automobiles destinées à des parcs de véhicules, nous définissons un classement basé sur le nombre total d'auto- mobiles que possède et utilise chaque consommateur (c.-à-d. la taille du parc d'automobiles). Dans la section 5, nous examinons plus en détail le choix de f .

Tableau 2
Estimations directes d'enquête, erreurs types et coefficients de variation pour certains produits

Produit (k)	Estimation du nombre de consommateurs		Estimation du nombre d'automobiles	
	Total	Y_k^*	Total	R_k

A	59,890	.3843	270,051	.3781
	(2,651)	(.0144)	(35,704)	(.0315)
	(.044)	(.037)	(.132)	(.083)
B	34,282	.2200	153,518	.2149
	(1,960)	(.0117)	(8,653)	(.0131)
	(.057)	(.053)	(.056)	(.061)
C	23,363	.1499	81,381	.1139
	(1,602)	(.0098)	(17,559)	(.0194)
	(.069)	(.065)	(.216)	(.170)
	13,857	.0889	25,312	.0354
	(1,311)	(.0081)	(2,906)	(.0039)
	(.095)	(.091)	(.115)	(.110)
E	9,025	.0579	24,370	.0341
	(1,146)	(.0072)	(7,336)	(.0101)
	(.127)	(.124)	(.301)	(.296)
F	5,125	.0329	13,724	.0192
	(676)	(.0043)	(2,369)	(.0030)
	(.132)	(.131)	(.173)	(.156)
G	7,518	.0482	11,031	.0154
	(1,015)	(.0064)	(1,456)	(.0022)
	(.135)	(.133)	(.132)	(.143)

modèle, produit par le fabricant A, qui occupe une part appréciable du marché des parcs de voitures. Les autres produits du tableau 2 occupent des petites parts de marché. Les produits F et G sont plutôt destinés au marché des voitures de fonction. La liste du tableau 2 est incomplète, ce qui explique que la somme des parts de marché n'est pas égale à un. Il convient aussi de souligner que les caté- gories de produits ne s'excluent pas mutuellement. De façon générale, on juge que l'enquête donne des résultats satisfaisants mais on s'aperçoit que les estimations relatives aux fabricants ou aux modèles qui ont de faibles parts de marché sont instables sur plusieurs années. Cela peut se remarquer par le coefficient de variation, qui est supérieur à 0.1 pour des produits qui occupent une part de marché modeste et qui peut dépasser 0.15 ou 0.2 dans certains cas. Cette instabilité se répercute sur les estimations de la variance de même que les estimations des ventes totales ou des parts de marché pour les produits.

4. MÉTHODE FONDÉE SUR UN MODÈLE

Étant donné le plan de sondage, on ne peut s'attendre à améliorer l'efficacité des estimateurs directs dans le cadre classique des enquêtes par sondage. La façon habituelle de procéder est d'utiliser de l'information supplémentaire

comprend les voitures achetées à l'intention des représen-
tants, lesquelles sont en grand nombre. Elle comprend
aussi les voitures de luxe achetées à l'intention des diri-
geants et des autres membres de la haute direction de
grandes entreprises, de même que les voitures achetées par
de petites sociétés, comme des groupes de médecins, ou
des travailleurs indépendants, comme des propriétaires
d'usine. La population d'entreprises acheteuses – appelées
consommateurs – comprend donc un grand nombre de
petites entreprises qui n'achètent qu'une ou deux automo-
biles à quelques années d'intervalle.

Nous définissons Y_{ki} comme le nombre d'automobiles
de type k achetées par le consommateur i dans la période
de référence d'un an. Le type de produit k (c.-à-d. le
domaine) peut désigner un modèle particulier ou tous les
modèles d'un fabricant. Par conséquent, $Y_k = \sum_i Y_{ki}$ est
le nombre total d'automobiles de type k achetées par
l'ensemble des consommateurs. Soit Z_i le nombre total
d'automobiles de toutes sortes achetées par le consom-
mateur i et $Z = \sum_i Z_i$, le nombre total d'automobiles
vendues. La part de marché pour le produit de type k est
définie par l'équation $R_k = Y_k/Z$.

Nous posons aussi les relations suivantes

$$Y_{ki} = 1 \quad \text{si} \quad Y_{ki} > 0 \\ = 0 \quad \text{si} \quad Y_{ki} = 0$$

$$Z'_i = 1 \quad \text{si} \quad Z_i > 0$$

$$= 0 \quad \text{si} \quad Z_i = 0.$$

Ainsi, Y_{ki} et Z'_i sont des variables indicatrices pour les
consommateurs qui achètent respectivement des produits
de type k et au moins une automobile de n importe quelle
sorte dans la période de référence. Le nombre de consom-
mateurs qui achètent des produits de type k est donc défini
par l'équation $Y_k = \sum_i Y_{ki}$ et le nombre total de consom-
mateurs qui achètent au moins une voiture de n importe
quel type est défini par l'équation $Z' = \sum_i Z'_i$. Le taux
de pénétration du marché pour le produit de type k
– c'est-à-dire la proportion de consommateurs qui, parmi
ceux qui achètent une voiture de n importe quel type dans
la période de référence, achètent une voiture de type k –
est défini par l'équation $R_k = Y_k/Z'$.

Les quatre paramètres Y_k , R_k , Y'_k et R'_k peuvent tous
faire l'objet d'inférence dans une étude de marché et sont
définis comme des paramètres de population finie, notam-
ment des totaux de domaine ou des rapports de totaux de
domaine.

3. PLAN DE SONDAGE ET ESTIMATEURS DIRECTS

Le plan de sondage repose sur deux bases s'excluant
mutuellement et peut être considéré comme un plan
d'échantillonnage simple stratifié comportant dix strates.

La première base de sondage consiste dans un registre

(Dun et Bradstreet) de 35,000 compagnies réparties dans
huit strates selon le nombre d'employés et la nature des
activités de l'entreprise (fabrication ou distribution). La
deuxième base de sondage consiste dans un vaste registre
de 1,4 million d'abonnés au service commercial de British
Telecom, répartis dans deux strates: lignes privées et lignes
commerciales. Notons qu'il s'agit d'abonnés au service
commercial dans les deux cas; une ligne commerciale est
attribuée si l'abonné dispose de plusieurs locaux.

À l'aide de données d'enquête antérieures, on a réparti
l'échantillon de façon optimale au moyen de la formule
de Neyman afin de réduire au maximum la variance de
l'estimateur du nombre total d'automobiles achetées (Z).
On a recueilli des données sur les achats d'automobiles
immédiatement après la fin de l'année de référence. Les
tailles de strate $\{N_h\}$ et les tailles d'échantillon $\{n_h\}$ pour
les strates $h = 1, \dots, 10$ figurent dans le tableau 1.

Tableau 1

Base de sondage: effectif et poids par strate				
Strate (h)	Taille de	Effectif	Poids	
	la strate N_h	n_h	$\pi_h^{-1} = N_h/n_h$	
British Telecom:				
Ligne privée	389,445	1,150	338,65	
Ligne commerciale	1,007,399	7,406	136,02	
Dun et Bradstreet:				
Fabrication	50-99 employés	6,646	235	28,28
	100-499	6,826	1,113	6,13
	500-999	992	520	1,91
	1,000 +	1,110	849	1,31
Distribution	50-99 employés	8,703	472	18,44
	100-499	7,625	1,437	5,31
	500-999	1,133	484	2,34
	1,000 +	1,523	1,117	1,36
Total	1,431,402	14,783		96,83

L'échantillon est un échantillon simple stratifié réparti
de façon non proportionnelle et les estimateurs directs de
même que les variances correspondantes sont connus avec
exactitude. La stratification donne des poids d'échantil-
lonnage très différents (de 1,31 à 338,65); elle est utile mais
loin d'être idéale. Beaucoup de consommateurs n'achètent
aucune automobile dans l'année de référence, de sorte que
chaque strate renferme un mélange de valeurs nulles et de
valeurs non nulles. La proportion de valeurs nulles dans
chaque strate est évidemment élevée pour un produit de
type k quelconque.

Le tableau 2 contient les estimations directes d'enquête,
les erreurs types estimées (voir Holt et Holmes (1993) pour
la façon dont elles ont été calculées) et les coefficients de
variation pour certains produits de fabricants d'automobi-
les. Les produits A et B représentent tous les modèles de
deux grands fabricants. Le produit C représente un seul

Estimation pour petits domaines dans des plans de sondage avec probabilités inégales

D. HOLT et D.J. HOLMES¹

RÉSUMÉ

L'estimation de totaux et de moyennes pour un domaine à l'aide de données d'échantillon est un problème courant. Lorsque le domaine est grand, l'échantillon observé est généralement suffisamment grand pour que des estimateurs directs, fondés sur un plan, soient assez précis. Mais lorsque le domaine est petit, la taille de l'échantillon observé est faible et les estimateurs directs ne conviennent plus. L'estimation pour petite région est un cas typique et des solutions de rechange comme l'estimation synthétique et l'estimation fondée sur un modèle ont été élaborées pour contourner le problème. Ces deux méthodes ont ceci de caractéristique, que de l'information est "empruntée" à d'autres petits domaines (ou petites régions) de manière à obtenir des estimateurs de paramètres plus précis, lesquels sont ensuite combinés avec de l'information supplémentaire, comme des moyennes ou des totaux pour population, tirée de chaque petite région afin d'établir une estimation plus précise de la moyenne ou du total pour un domaine (ou une région). Dans cet article, nous étudions un cas d'échantillonnage avec probabilités inégales où il n'existe pas d'information supplémentaire et où l'emprunt d'information n'est pas permis; pourtant, des estimateurs simples fondés sur un modèle sont élaborés dans ces conditions et semblent offrir des gains d'efficacité appréciables. L'exemple que nous utilisons ici est une étude de marché mais les domaines d'application de ces méthodes sont nombreux.

MOTS CLÉS: Estimation synthétique; estimation fondée sur un plan; estimation pour petites régions; estimation fondée sur un modèle; parts de marché.

1. INTRODUCTION

Dans cet article, nous nous intéressons à l'estimation de totaux et de moyennes pour un domaine à partir d'un échantillon réparti de façon non proportionnelle. Certains domaines peuvent être grands, auquel cas la taille effective de l'échantillon peut aussi être élevée et des estimateurs fondés sur un plan (ou estimateurs directs) conviendront. D'autres domaines peuvent être petits, auquel cas la taille effective de l'échantillon sera également faible et les estimateurs fondés sur un plan (ou estimateurs directs) seront trop imprécis pour être utilisés convenablement. Nous allons illustrer les méthodes proposées à l'aide d'un exemple où sont estimées des ventes, des parts de marché et des taux de pénétration pour des produits dans une étude de marché. Les domaines sont constitués de fabricants d'automobiles ou de modèles de voiture. Cependant, la méthode générale peut s'appliquer à d'autres échantillons d'entreprises ou d'institutions répartis de façon non proportionnelle.

Le problème étudié ressemble au problème de l'utilisation de l'estimation synthétique pour de petites régions (Gonzales 1973; Gonzales et Hoza 1978; Plarek et coll. 1987). L'estimation synthétique dépend normalement de deux facteurs: i) l'emploi combiné de variables auxiliaires et de moyennes ou de totaux de population pour chaque petite région (ou petit domaine) en vue d'améliorer les estimations par la stratification a posteriori ou l'estimation par régression, et ii) l'amélioration des estimations par regroupement de données de plusieurs petites régions

2. EXEMPLE: ÉTUDE DE MARCHÉ

(ou petits domaines). Dans le cas qui nous occupe, il n'existe pas de moyennes ou de totaux de population et comme l'objectif essentiel est de comparer des domaines (c.-à-d. fabricants et produits particuliers), l'emprunt d'information de l'un à l'autre est inconcevable. Nous posons une catégorie d'estimateurs synthétiques qui n'utilisent aucune de ces deux méthodes et qui, pourtant, sont préférés aux estimateurs directs. Les estimateurs proposés ont une structure simple et une interprétation intéressante et peuvent être justifiés suivant un ensemble d'hypothèses de modèle qui peuvent être testées selon l'hypothèse générale des plans de sondage non informatifs.

Les responsables d'études de marché estiment souvent le volume total des ventes et la part de marché pour chaque fabricant d'un produit. Nous examinons le cas des automobiles achetées au cours d'une année en vue d'être utilisées comme voitures d'entreprise. Il faut des estimations de totaux et de parts de marché pour chaque fabricant d'automobiles et pour les modèles qui constituent le plus souvent les parcs de véhicules.

Les termes "parc" et "entreprise" ont chacun une définition assez large. Une voiture d'entreprise est définie comme une automobile qui est achetée à des fins commerciales plutôt que pour un usage privé et qui est associée à une entreprise au sens le plus large. Cette définition

¹ D. Holt et D.J. Holmes, Department of Social Statistics, University of Southampton, Highfield, Southampton, UK, SO95NH.

une valeur n_{\min}^* , soit l'effectif qui permet de produire des estimations directes tout juste acceptables. Notons toutefois que n_{\min}^* , tel qu'il est défini ici, est dépendant des caractéristiques.

Dans ses commentaires, Fuller décrit sommairement une méthode d'estimation pour petites régions qui exploite un modèle fondé sur les composantes de la variance et qui, pourtant, utilise des poids fixes pour garantir la cohérence interne des estimations relatives à diverses caractéristiques. Outre la cohérence interne des estimations régionales, il faut parfois que la somme des totaux estimés des petites régions qui forment une grande région corresponde à l'estimation directe officielle pour la grande région. Une façon d'assurer cette concordance est d'ajuster les estimations régionales à l'estimation directe pour la grande région au moyen, par exemple, d'un simple coefficient de correction; cependant, si les coefficients de correction dépendent des caractéristiques, la cohérence interne mentionnée précédemment n'existera plus. On peut garantir simultanément les deux types de cohérence à la condition que l'estimateur direct pour la grande région soit un estimateur par régression généralisé, $X_e + (X_e - X_e^*)\beta$, et qu'on utilise l'estimateur direct modifié $X_{seg,a}^{seg,a} = X_{e,a} + (X_a - X_{e,a})\beta$ (section 6.1 de notre article) pour les petites régions.

Comme le mentionne Fuller, il est possible d'estimer la valeur moyenne du biais au carré d'un estimateur pour un sous-ensemble quelconque de petites régions. Nous aimerions préciser de nouveau ici que la valeur moyenne du biais pour un ensemble de petites régions n'a pas de rapport direct avec une petite région en particulier. C'est pourquoi nous préférons utiliser, quand c'est possible, des estimateurs qui sont approximativement non biaisés selon le plan. Si nous n'avons d'autre choix que d'utiliser un estimateur fondé sur un modèle, nous devons nous efforcer de trouver des covariables pertinentes pour lesquelles il existe de l'information supplémentaire fiable afin de réduire au minimum le biais résiduel de l'estimateur.

Peut-être à cause des délais qu'implique inévitablement la production des données du recensement, aucun des deux

critiques ne considère le recensement comme une source de données pour de petits domaines. À cet égard, il convient de mentionner qu'on envisage actuellement une formule qui prévoirait une sorte de grande enquête postcensitaire continue en remplaçant le questionnaire détaillé utilisé dans le recensement décennal. Cette formule, appelée méthode des échantillons avec renouvellement complet, est décrite dans Kish (1990); une formule semblable, appelée méthode de dénombrement continu, est décrite dans Alexander (1994). Cette méthode comporte un certain nombre de fonctions qui méritent d'être examinées dans la mesure où elles peuvent devenir autant de moyens de produire à meilleur marché et dans de courts délais des données pour petits domaines.

Enfin, nous tenons à préciser qu'en soulignant l'importance de considérer l'estimation pour domaine à l'étape de la conception du plan de sondage, surtout en ce qui a trait aux domaines de taille moyenne, nous ne voulons aucune-ment minimiser le rôle des modèles dans l'estimation qui s'applique à de très petits domaines.

Nous espérons que l'approche générale présentée dans l'article, de même que les commentaires perspicaces des critiques, en particulier les conseils et les mises en garde formulés par Kalton dans le dernier paragraphe de ses commentaires, permettront aux concepteurs de plans de sondage et aux chercheurs de trouver les solutions qui conviennent pour les problèmes qui les intéressent.

SOURCES ADDITIONNELLES

ALEXANDER, C.H. (1994). A prototype continuous measurement system for the U.S. Census of Population and Housing. Population Association of America, Miami, Florida, 5 mai, 1994.

KISH, L. (1990). Recensement par étapes et échantillons avec renouvellement complet. *Techniques d'enquête*, 16, 67-77.

RÉPONSE DES AUTEURS

on recourt au plan avec formation de grappes pour réduire les coûts de l'enquête. Compte tenu de l'évolution actuelle des méthodes de collecte de données – par exemple l'usage moins fréquent de l'interview sur place au profit de l'interview assistée par ordinateur – il est devenu nécessaire de revoir périodiquement les modèles coût-variance qui conditionnent les décisions concernant la formation de grappes. Une autre question que nous n'avons pas traitée dans l'article est celle de l'effet du renouvellement de l'échantillon dans les enquêtes permanentes. Pour une période donnée, il se peut que certains petits domaines n'aient pas un effectif suffisant pour qu'on puisse établir des estimations fiables. Cependant, à mesure que les unités de l'échantillon se succèdent par renouvellement, la taille cumulée ou effective de l'échantillon dans les domaines augmente et on peut ainsi calculer, par domaine, des estimations fiables, quoique biaisées chronologiquement. Par un choix judicieux de plans de renouvellement, les concepteurs d'enquêtes peuvent maximiser la taille cumulée de l'échantillon dans une période donnée. Par exemple, pour des estimations trimestrielles d'une enquête mensuelle, le plan de renouvellement optimal est $[1(2)]^k$, c'est-à-dire la répétition de k fois la séquence "un mois dans l'échantillon, deux mois hors de l'échantillon". Ce raisonnement s'applique à celui de Leslie Kish sur le cumul d'échantillons dans le temps.

Kalton nuance la mise en garde faite sur l'emploi des estimations indirectes en proposant une erreur quadratique moyenne pondérée, par laquelle on associe un poids plus grand que 1 au terme du biais pour tenir compte de ce que le biais de l'estimateur indirect peut être plus élevé que le biais de l'estimateur direct. Deux raisons font que le biais peut être plus élevé que le laisse croire le modèle pour les effets de petite région: la variation aléatoire à l'intérieur du modèle et la débiaillance du modèle. Il est utile de se rappeler ici la suggestion de Fay et Herriot (1979), qui proposent qu'il y ait un écart de moins d'un erreur type entre une estimation combinée et une estimation de plan; cette mesure permet de concevoir la possibilité que l'estimateur de modèle renferme un biais élevé quel qu'en soit la raison. Kalton affirme aussi, comme nous, qu'étant donné un estimateur direct de qualité acceptable, on pourra choisir en pratique d'utiliser cet estimateur même si l'erreur quadratique moyenne estimée correspondante est plus grande que celle d'estimateurs fondés sur un modèle. Compte tenu de la possibilité toujours réelle d'une débiaillance du modèle, on favorisera cette approche du "meux vaut prévenir que guérir", du moins tant qu'on n'aura pas expérimenté certains estimateurs indirects dans des conditions particulières. Et cela n'empêche pas qu'il y ait des situations où il faut faire preuve d'un peu d'audace.

Dans ses observations sur l'estimateur tributaire de la taille de l'échantillon, Kalton laisse entendre que l'attribution d'un poids nul à la composante synthétique peut poser un problème si la taille réelle de l'échantillon pour un petit domaine excède la taille prévue, car celle-ci peut n'être pas assez élevée pour qu'on puisse produire des estimations directes acceptables. Une solution possible serait d'utiliser

Nous remercions Wayne Fuller et Graham Kalton pour leurs commentaires stimulants qui complètent très bien les idées que nous avons exprimées dans l'article. Plusieurs de ces commentaires apportent des éclaircissements sur certains points et renforcent les arguments exposés. Encouragés par cet appui, nous aimerions approfondir certains points touchant le plan de sondage et, en même temps, répondre aux principaux commentaires des critiques.

Il est clair que le concepteur de plans de sondage s'efforce, malgré des contraintes opérationnelles, d'optimiser le plan dans le but de respecter les objectifs d'une enquête. Les grandes enquêtes poursuivent habituellement plusieurs objectifs, et il est fort probable que le concepteur a peu de prise sur l'établissement d'un ordre de priorité pour ces objectifs. Néanmoins, c'est justement à l'étape de l'établissement d'un ordre de priorité qu'il faut faire ressortir avec vigueur les besoins en données régionales, particulièrement en ce qui concerne les grandes enquêtes permanentes.

Dans les années soixante et soixante-dix, la plupart des pays ont mis l'accent sur les estimations infranationales (État, province) et modifié dans une certaine mesure les plans de sondage antérieurs qui, eux, visaient à optimiser les estimations nationales. Par exemple, on a utilisé des fractions de sondage différentes pour être sûr que les petits États ou les petites provinces contiendraient au moins tel nombre d'unités de l'échantillon. Vu la demande croissante de données au niveau infraprovincial ou infra-État (comité, district et municipalité), il faut modifier davantage les plans antérieurs, qui, eux, prévoient une répartition optimale de l'échantillon au niveau national. Cette modification peut consister à utiliser des fractions de sondage différentes pour les divisions administratives d'un État ou d'une province. Par exemple, si l'on souhaite produire des estimations infraprovinciales de qualité comparable, on attribuera sans doute à chaque province un échantillon qui sera à peu près proportionnel au nombre de régions infraprovinciales que contient la province. Cela n'équivaut peut-être pas à la taille relative de la population des provinces. Comme nous l'avons mentionné dans la section 5.4 de notre article, il faudrait privilégier une approche ascendante pour la répartition de l'échantillon. Les pertes enregistrées aux niveaux d'aggrégation supérieurs, comme les gains aux niveaux inférieurs, varieraient selon les enquêtes, mais il est probable que, dans beaucoup de cas, une légère augmentation du coefficient de variation au niveau national se traduirait par des gains appréciables au niveau des petites régions.

Kalton souligne l'importance de limiter la grosseur des grappes pour l'estimation de la variance; il est avantageux d'accroître le nombre de degrés de liberté en formant un grand nombre de petites grappes au lieu d'un petit nombre de grosses grappes. Précisons que l'agglomération d'unités d'échantillonage présente un autre inconvénient du point de vue de l'estimation, en particulier de l'estimation pour petites régions; en effet, un plan qui prévoit la formation de grosses grappes produira des effets de plan prononcés, même pour de petits domaines planifiés. En règle générale,

l'estimateur direct ou de l'estimateur indirect. Or, d'après le raisonnement ci-dessus, on pourrait tout aussi bien minimiser le poids de l'estimateur indirect, à la condition que l'estimateur combiné soit suffisamment précis. On pourrait aussi déterminer les poids selon une valeur probable maximum de l'EQM, plutôt que selon l'EQM prévue, afin de réduire la probabilité d'un biais important dans l'estimateur combiné.

Je ne comprends pas très bien l'explication des estimateurs tributaires de la taille d'échantillon définis par les équations (6.11) et (6.12) de SGM mais suivant certaines hypothèses, ces estimateurs pourraient cadrer dans l'argumentation présentée ci-dessus. Étant donné un plan d'échantillonnage avec probabilités égales et une valeur $\delta = 1$, ces estimateurs se ramènent à l'estimateur direct lorsque la taille effective de l'échantillon est égale ou supérieure à la taille prévue. Si on suppose que la taille prévue de l'échantillon implique un degré de précision acceptable pour une petite région, l'argumentation ci-dessus est respectée. Si la taille effective de l'échantillon est plus petite que prévu, l'estimateur tributaire de la taille d'échantillon devient la moyenne pondérée d'un estimateur direct et d'un estimateur indirect. Si on suppose que la taille prévue de l'échantillon représente l'effectif minimum nécessaire pour obtenir le degré de précision voulu, l'argumentation ci-dessus est de nouveau respectée. Si les observations précédentes constituent le fondement des estimateurs tributaires de la taille d'échantillon, il semblerait utile d'étendre ces estimateurs aux situations où la taille prévue de l'échantillon ne correspond pas à celle permettant tout juste d'atteindre le degré de précision voulu.

Comme je l'ai mentionné plus haut, l'information supplémentaire compte pour beaucoup dans la production d'estimations régionales précises. Cette information peut servir à accroître la précision des estimations fondées sur un plan ou elle peut être utilisée dans les modèles auxquels on recourt par l'approche dépendante de modèles. Idéalement, il faut de l'information supplémentaire qui soit étroitement liée aux variables étudiées dans l'enquête. En extrayant régulièrement de sources administratives ou autres des données auxiliaires récentes concernant les petites régions, on peut contribuer à améliorer les données régionales.

Bien que les auteurs évoquent la question plus générale des petits domaines, ils concentrent leur attention sur les petites régions. D'ailleurs, la question des petites régions est celle qui est le plus souvent traitée dans la littérature statistique et qui est en rapport avec l'application de méthodes d'estimation indirecte. Cela peut s'expliquer d'une part, par le fait que les domaines socio-économiques et les autres petits domaines pertinents (ex.: groupes d'âge-sexe) sont habituellement assez peu nombreux comparativement aux petites régions de sorte qu'on peut concevoir

BIBLIOGRAPHIE

- pour ces domaines un plan de sondage qui produise des estimations d'une précision acceptable pour chacun d'eux et d'autre part, par le fait que les domaines socio-économiques et démographiques sont souvent définis en fonction de la possibilité d'obtenir des estimations de plan d'une précision acceptable pour ces domaines (par ex., utiliser des groupes d'âge plus étendus pour certains domaines); toutefois, en ce qui concerne les domaines spatiaux, les territoires sont définis à l'avance et aucun regroupement n'est possible. Il en est ainsi d'une part, parce que les données auxiliaires sont en nombre insuffisant pour être utilisées dans les modèles statistiques s'appliquant aux domaines de ce genre et d'autre part, parce que l'analyse des domaines socio-économiques a souvent pour but d'établir des comparaisons entre les domaines. Or, ces comparaisons sont faussées si l'estimation relative à un domaine est "renforcée" par de l'information provenant d'autres domaines (voir, par exemple, Schabale 1992). Cela nous amène à dire qu'on ne devrait pas utiliser les estimations indirectes sans faire preuve d'esprit critique.
- En conclusion, je suis d'accord avec l'approche générale présentée dans cet article. Lorsque c'est possible, les échantillons devraient être conçus de manière à produire des estimations directes pour petite région d'une précision acceptable, et les plans de sondage devraient être élaborés en ce sens. De même, lorsque c'est possible, on devrait se servir de données auxiliaires pour accroître la précision des estimations directes pour petite région. Lorsqu'il faut recourir à des estimations indirectes, on doit procéder avec prudence. Il faut élaborer soigneusement les modèles, chercher des estimateurs qui résistent bien aux défaillances d'un modèle et faire des analyses pour évaluer la pertinence des estimations indirectes. En l'absence de bons indices pour mesurer la qualité des estimations indirectes, il est essentiel de distinguer clairement ces estimations des estimations fondées sur un plan. Comme les estimations indirectes ne conviennent pas à tous les usages, l'utilisateur doit s'assurer que les estimations indirectes dont il veut se servir répondront à ses besoins.
- SÄRNDAAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SCHABALE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings of the International Scientific Conference on Small Area Statistics and Survey Design*, (Vol. 1), 95-114, Central Statistical Office of Poland, Warsaw.

au moyen de l'échantillon. Dans ces circonstances, l'estimation de X_0 est convergente selon le plan, peu importe le modèle utilisé (Särndal 1984). En outre, il est possible d'estimer les valeurs de B à l'aide des données de l'échantillon relatives au domaine étudié seulement, ce qui donne ce que SGM appellent un estimateur direct, ou à l'aide des données de l'échantillon global, ce qui donne un estimateur direct modifié. Un facteur déterminant dans le choix entre l'estimateur direct et l'estimateur direct modifié dans ces circonstances est de savoir si les valeurs de B pour l'échantillon global valent aussi pour le domaine. Si ce n'est pas le cas, il faut ajouter dans le modèle pour l'échantillon global des termes d'interaction entre X et les indicateurs de domaine. Avec un ensemble complet de termes d'interaction, l'estimateur direct modifié se ramène en fait à l'estimateur direct ordinaire.

On doit recourir à une approche dépendante d'un modèle lorsque les estimations fondées sur un plan ne sont pas assez précises, malgré qu'on ait utilisé le plus efficacement possible l'information supplémentaire qui était disponible. Il arrive en effet qu'on ne soit pas capable de calculer une estimation directe à cause de l'absence d'éléments de l'échantillon dans la petite région. Dans ces circonstances, on doit recourir à un modèle statistique afin d'"emprunter de l'information" à d'autres régions. Ces modèles reposent sur des hypothèses (ex.: $E_a = 0$ dans l'exemple ci-dessus) et la qualité des estimations obtenues à l'aide de ces modèles dépend de la validité des hypothèses. Comme une hypothèse est toujours un peu inexacte, les estimations régionales sont entachées d'un biais, et comme il s'agit d'estimations indirectes biaisées, on se sert largement de l'erreur quadratique moyenne (EQM) de plan pour en mesurer la qualité (notons que $EQM = V' + B^2$ où V' est la variance et B est le biais de l'estimation).

La manière la plus courante de comparer la qualité d'une estimation directe avec celle d'une estimation indirecte est de comparer la variance (V) de la première avec l'EQM de la seconde. Or, en lisant l'article de SGM, j'en suis venu à me demander si l'EQM était la bonne mesure pour évaluer la qualité d'un estimateur indirect. Dans la pratique, il est possible d'estimer la variance V d'un estimateur direct, ce qui n'est pas le cas pour l'EQM de plan d'un estimateur indirect. Compte tenu de cette observation, on préférera nettement l'estimateur direct si $V = EQM$. De fait, on optera le plus souvent pour l'estimateur direct de celui-ci offre un degré de précision acceptable, peu importe le rapport probable des valeurs V et EQM. Dans d'autres cas, si B est le biais prévu, on pourra préférer l'estimateur direct à l'estimateur indirect, à moins que $V > V' + KB^2$, où K est un facteur plus grand que 1 qui tient compte de la possibilité que le biais inconnu soit plus élevé que prévu.

On peut appliquer le même raisonnement aux estimateurs combinés (ou composites), qui sont la moyenne pondérée d'un estimateur direct et d'un estimateur indirect. On considère souvent que la meilleure manière de déterminer les poids est de minimiser l'erreur quadratique moyenne de l'estimateur combiné, ce qui donne des poids inversément proportionnels à V et à EQM, selon qu'il s'agit de

petite région étant donné un plan avec grappes, puisque les UPÉ tirées dans chaque petite région seront vraisemblablement peu nombreuses. Une estimation de la variance fondée sur les UPÉ d'une petite région sera donc imprécise, étant donné quelques degrés de liberté, et on préférera probablement une fonction de variance généralisée (en supposant, par exemple, que l'effet de plan au niveau national vaut pour chaque petite région). Autrement dit, même si l'estimation proprement dite est fondée sur un plan, l'estimation de la variance correspondante peut être une estimation indirecte, renforcée par les données d'autres régions. C'est pourquoi il est préférable d'avoir un plan de sondage qui prévoit le moins de grappes possible, même lorsqu'il s'agit de petits domaines planifiés. Toutefois, il est plus important de penser à modéliser les estimations proprement dites que de se concentrer sur la modélisation des variances.

Un des principes fondamentaux de l'approche fondée sur un plan est que l'information supplémentaire qui existe sur la population peut servir à l'étape de la conception du plan de sondage, à celle de l'analyse ou aux deux à la fois. Lorsqu'il existe de l'information sur des variables auxiliaires qui ont un rapport étroit avec la variable étudiée, on peut réaliser des gains appréciables en précision. L'utilisation d'information supplémentaire à l'étape de l'analyse, par l'intermédiaire de techniques comme la stratification a posteriori, l'estimation par quotient ou par régression et l'estimation de différences, a un attrait particulier en ce qui concerne l'estimation pour petite région. Il convient de souligner que les estimateurs par quotient et par régression peuvent trouver leur justification dans des hypothèses portant sur le modèle qui met en relation la variable étudiée (Y) et les variables auxiliaires (X), mais les estimateurs seront convergents selon le plan peu importe si le modèle convient ou non. L'utilisation d'un modèle approprié permet de réaliser les meilleurs gains en précision mais quel que soit le modèle choisi, les estimations sont approximativement non biaisées. On peut illustrer cela par un exemple simple, où la valeur des variables X_1, X_2, \dots, X_p est connue pour chaque élément de la population et la combinaison linéaire $Y_i = B_0 + B_1X_{1i} + \dots + B_pX_{pi}$ est utilisée pour estimer Y_i , la valeur de la variable Y pour l'élément i de la population. Supposons pour des raisons de simplicité que les valeurs de B sont déterminées à l'aide de données externes, indépendantes de l'échantillon. Étant donné l'équation $Y_i = X_i + e_i$, le total de domaine est $Y_a = \sum_{i \in a} Y_i + \sum_{i \in a} e_i = Y_a + E_a$. Comme Y_a est connu, le problème revient à estimer E_a . À partir d'un échantillon d'éléments du domaine a , on peut estimer E_a au moyen de l'équation $E_a = \sum_{j \in a} e_j / \pi_j$, où π_j est la probabilité de sélection pour l'élément j de l'échantillon. L'estimateur E_a est non biaisé, peu importe si le modèle utilisé est valide ou non. En fait, la procédure est conçue de telle manière qu'au lieu d'estimer directement Y_a , on estime E_a et on ajoute à cette valeur une constante connue Y_a . Pour que la procédure soit efficace, la variance des e_j pour le domaine doit être inférieure à celle des Y_j . Il n'est pas nécessaire que $E_a = 0$. Le raisonnement est le même dans le cas plus habituel où les valeurs de B sont estimées

COMMENTAIRES

GRAHAM KALTON¹

les territoires de bureaux de placement. Dans de telles circonstances, on peut devoir remplacer les méthodes classiques d'inférence fondée sur un plan par des méthodes à modèles, qui utilisent pour l'estimation un modèle statistique par lequel on renforce les estimations d'une petite région par des données recueillies dans l'enquête. On peut devoir recourir aussi aux méthodes fondées sur un modèle pour de petits domaines non planifiés, pour lesquels on n'avait pas prévu la nécessité d'un suréchantillonnage à l'étape de la conception.

Compte tenu de l'intérêt pour les estimations régionales, un nombre appréciable d'ouvrages sont parus sur les méthodes d'estimation pour petit domaine fondées sur un modèle. Cependant, peu d'ouvrages ont été écrits jusqu'à maintenant sur les questions que traite l'article de SGM, lesquelles méritent une plus grande attention. Comme les auteurs, je crois qu'il faut envisager la question de l'utilisation d'estimateurs pour petite région fondés sur un modèle avec circonspection. C'est pourquoi je suis heureux de voir que SGM examine les méthodes qui permettent d'établir des estimations régionales au moyen de l'inférence fondée sur un plan.

À mon point de vue, la première étape dans le calcul d'estimations régionales est de vérifier si on peut obtenir des estimations ayant un degré de précision acceptable avec des méthodes fondées sur un plan. Si les domaines sont définis à l'avance, on devrait envisager de concevoir le plan d'échantillonnage de manière à faciliter la production d'estimations régionales. Cela peut signifier de faire en sorte que les petites régions ne chevauchent pas de strates et qu'il y ait un échantillon de taille suffisamment grande pour chaque petite région. Une autre solution proposée par SGM consiste à réduire au maximum la formation de grappes. Plus la formation de grappes est limitée, moins la taille de l'échantillon dans chaque petite région est laissée au hasard. À cet égard, le principal avantage de la limitation de la formation de grappes est de permettre la production d'estimations pour de petites régions qui n'étaient pas définies à l'étape de la conception du plan de sondage. Lorsque de petites régions pour lesquelles des estimations sont prévues ne chevauchent pas des strates, la taille de l'échantillon dans chaque petite région devrait être contrôlée raisonnablement même lorsqu'il s'agit d'un échantillon en grappes (pourvu que les mesures de taille utilisées dans l'échantillonnage avec PPT soient raisonnables). Or, même dans le cas d'estimations prévues, il arrivera souvent qu'on se demande comment calculer les estimations de la variance pour une

Comme le mentionnent Singh, Gambino et Mantel (SGM), on cherche de plus en plus à réaliser des enquêtes qui vont produire des estimations pour des domaines de tous genres et de toutes tailles. Cette tendance est observée dans de nombreux pays dans le monde. D'une part, elle peut être simplement le résultat d'une évolution naturelle des exigences des analystes d'enquête, qui auparavant se contentaient d'estimations nationales et d'estimations touchant un petit nombre de domaines majeurs mais qui, aujourd'hui, veulent comparer des estimations relatives à de nombreux types de domaines différents. D'autre part, cette tendance vient des besoins des responsables de l'action gouvernementale, qui recherchent de l'information pour domaine afin d'analyser les effets des programmes courants sur différents domaines, de prévoir les conséquences d'une révision des politiques et de suivre l'exécution des programmes. Les données relatives à des divisions administratives (par ex., province ou État, comté, arrondissement scolaire) présentent un intérêt particulier pour l'élaboration de programmes (par ex., lorsqu'il s'agit de définir les régions à faible revenu aux fins de l'élaboration des programmes de soutien public).

Dans certains cas, on peut réussir à obtenir des estimations pour domaine suffisamment précises à l'aide des méthodes d'inférence fondée sur un plan qui sont utilisées habituellement dans l'analyse de données d'enquête. C'est le cas notamment pour les grands domaines, où la taille des échantillons est suffisamment grande pour qu'on ait le degré de précision voulu. Cela est possible aussi pour de petits domaines, pourvu qu'ils soient définis à l'avance et que le plan de sondage soit élaboré de manière à produire des échantillons de taille acceptable. Ainsi, aux États-Unis par exemple, la National Health and Nutrition Examination Survey et la Continuing Survey of Food Intakes by Individuals utilisent des fractions de sondage qui varient selon l'âge, le sexe ou l'origine ethnique dans le premier cas et selon l'âge et le sexe ou le niveau de faible revenu dans le second cas, afin de produire des échantillons acceptables pour les domaines créés par le recoupement de ces variables. La Current Population Survey des E.-U. utilise des fractions de sondage différentes selon les États afin de produire des estimations de l'emploi pour l'État. La faiblesse de cette méthode ressort clairement lorsqu'il y a un grand nombre de petits domaines, auquel cas la somme des effectifs de chaque domaine dans l'échantillon donne une taille d'échantillon global très élevée. On observe souvent cela pour les petites divisions administratives comme les comtés, les arrondissements scolaires et

¹ Graham Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

J'estime que dans la plupart des cas d'estimation pour domaine, les gains d'efficacité découlent surtout de l'utilisation judicieuse de l'information supplémentaire. Par conséquent, les efforts consacrés à l'obtention d'information supplémentaire de qualité sont justifiés. Si nous pouvons trouver une variable x qui soit étroitement corrélée avec la variable y , il n'y aura plus autant de variabilité à répartir entre la variance d'une région à l'autre et la variance d'échantillonnage.

REMERCIEMENTS

Je remercie Jay Breidt pour ses commentaires.

BIBLIOGRAPHIE

BATTESE, G.E., HARTER, R.M., et FULLER, W.A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

FAY, R.E. (1987). Application of multivariate regression to small domain estimation. Dans *Small Area Statistics*. (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley.

FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.

FULLER, W.A., et HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. Dans *Small Area Statistics*. (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh). New York: Wiley.

HARVILLE, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *The Annals of Statistics*, 4, 384-395.

GHOSH, M., et RAO, J.N.K. (1993). Small area estimation: An appraisal. Manuscrit non-publié. Carleton University, Ottawa, Ontario, Canada.

PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Dans leur analyse de modèles, les auteurs soulignent

l'importance de produire des estimateurs de la fiabilité des estimateurs pour petites régions. Ils laissent entendre que les principaux estimateurs de l'erreur quadratique moyenne pour des méthodes fondées sur un modèle sont des estimateurs de la valeur moyenne de l'erreur quadratique moyenne. Bien que ce soit vrai, il convient de mentionner que la méthode des "composantes de la variance" ne suppose pas que l'erreur quadratique moyenne est la même dans chaque domaine. En outre, dans une enquête classique, les estimateurs de l'erreur quadratique moyenne ne sont pas nécessairement les mêmes dans tous les domaines. Par exemple, un des termes de l'estimateur de l'erreur quadratique moyenne utilisé dans la méthode des "composantes de la variance" est l'estimateur de la variance de l'estimateur direct. La variance estimée de l'estimateur direct sera une fonction de l'effectif de l'échantillon contenu dans le domaine et pourra aussi être une fonction de la variance estimée (directement) de l'estimateur direct pour ce domaine. Voir Battese, Harter et Fuller (1988), Harville (1976), Prasad et Rao (1990) et Ghosh et Rao (1993).

Dans leur analyse de plans de sondage, les auteurs expliquent que la courbe de la fonction de variance est souvent "aplatie" dans le voisinage de la répartition optimale des unités entre les strates. Une légère redistribution de l'échantillon peut avoir pour effet d'accroître sensiblement l'efficacité des estimations pour domaine sans réduire de façon notable l'efficacité des estimations globales. Il en va de même de la combinaison d'un estimateur direct et d'un estimateur synthétique. Par conséquent, si quelque un a une idée assez juste de la composante de variance liée aux petites régions, soit grâce à une étude antérieure qui portait sur la même population ou grâce à une étude qui portait sur une population similaire, et s'il doit produire des estimations dans un délai très limité, il est raisonnable, dans les circonstances, d'attribuer des poids fixes pour former la combinaison linéaire. Il y aura vraisemblablement peu de perte d'efficacité et les exigences de programmation pour la construction d'estimateurs seront beaucoup moins grandes. L'estimateur synthétique est un estimateur indiqué dans les circonstances, et beaucoup de praticiens l'utilisent.

Les auteurs évoquent la question de la propriété de fermeture liée à la construction d'estimateurs pour petites régions. Comme ils l'affirment, si on utilise une méthode dépendante des données comme la méthode des composantes de la variance, on obtient, pour chaque variable dépendante, des estimations qui n'ont pas la propriété de fermeture. Dans les circonstances, on peut toujours recourir à des méthodes multidimensionnelles. Voir, par exemple, Fuller et Harter (1987) et Fay (1987). Une autre méthode proposée par Fuller (1990) consiste à construire des estimateurs de composantes de la variance pour un sous-ensemble limité de variables, puis de se servir des estimations correspondantes comme variables de contrôle dans une méthode de régression. Celle-ci produit des poids pour les observations individuelles. Une fois que les poids sont déterminés, on peut construire un nombre indéfini de tableaux de données et toutes les estimations ont la propriété de fermeture.

$$\begin{aligned} \text{EQM}\{h_{(3)yt} | n_t\} &= (1 + n^{-1}) \\ &\times \sum_k N^{-2} N_t^2 n_t^{-1} V\{Y_{tj} - \beta X_{tj} | \ell = t\} \\ &+ (\mu_{xt} - \mu_{xt}^2) V\{b\} \\ &+ [\mu_{yt} - \mu_{yt} - \beta(\mu_{xt} - \mu_{xt}^2)]^2 + O(n^{-2}), \end{aligned}$$

où $V\{b\} = E\{(b - \beta^2), V\{a_\ell | \ell = t\}$ est la variance de la variable a pour le domaine t ,

et

$$\begin{aligned} \beta &= \left[\sum_k N^{-1} N_t V\{X_{tj} | \ell = t\} \right]^{-1} \\ &\times \sum_k N^{-1} N_t C\{Y_{tj}, X_{tj} | \ell = t\}. \end{aligned}$$

L'estimateur $h_{(1)yt}$ utilise uniquement l'information contenue dans l'échantillon de n_t observations. Par conséquent, toutes les propriétés de l'estimateur sont fonction de n_t et des paramètres du domaine. Le biais de régression est d'ordre n_t^{-1} , comme la variance. L'estimateur $h_{(2)yt}$ utilise les moyennes de domaine, mais utilise aussi l'échantillon complet pour l'estimation du coefficient de régression. En conséquence, la variance fondamentale est aussi d'ordre n_t^{-1} et sera plus élevée que la variance fondamentale de $h_{(1)yt}$ dans les cas où $\beta_t \neq \beta$. Or, la contribution de deuxième ordre à la variance est d'ordre $n_t^{-1} n^{-1}$ pour $h_{(2)yt}$ et d'ordre n_t^{-2} pour $h_{(1)yt}$. De plus, le biais de régression pour $h_{(2)yt}$ est d'ordre n^{-1} . Si les domaines étaient des strates, on pourrait désigner $h_{(1)yt}$ comme l'estimateur par régression simple et $h_{(2)yt}$ comme l'estimateur par régression combiné.

L'estimateur $h_{(3)yt}$ est un estimateur synthétique qui a une variance d'ordre n^{-1} , contrairement aux deux premiers estimateurs, qui ont une variance d'ordre n_t^{-1} . Cependant, cette réduction de la variance a un inconvénient: le biais est d'ordre un. On aura un biais nul seulement si la droite de régression est la même pour le domaine et pour la population entière.

On peut estimer la moyenne des erreurs quadratiques moyennes des trois estimateurs pour n'importe quel sous-groupe de petites régions. Si les n_t sont faibles, les variances estimées ne donneront qu'une information restreinte pour faire la différence entre les estimateurs. En outre, il n'existe qu'un degré de liberté pour le carré du biais pour un domaine particulier. Cependant, si l'écart pour un domaine devait être élevé par rapport à l'erreur-type, il faudrait reconsidérer l'estimateur synthétique.

COMMENTAIRES

W.A. FULLER¹

Les auteurs méritent des félicitations pour avoir exposé avec brio les problèmes de conception de plan et d'estimation qui ont rapport aux domaines. Les auteurs examinent l'estimation pour domaines planifiés en particulier les cas où les éléments d'un domaine peuvent être identifiés dans la base de sondage de même que l'estimation pour domaines non planifiés, notamment les domaines dont les éléments ne peuvent être identifiés à l'aide de la base de sondage. Cet article est un élément positif supplémentaire dans la littérature de plus en plus abondante qui existe sur la question de l'estimation pour domaine.

Les auteurs font une description particulièrement bonne des opérations de planification, de collecte des données et de traitement qui sont exécutées dans le cadre des enquêtes de Statistique Canada. Ils mentionnent les problèmes qui sont liés depuis toujours à la conception des plans de sondage: arbitrage entre la nécessité de recourir à l'estimation pour domaine et le désir d'obtenir une certaine efficacité aux niveaux (d'agrégation) supérieurs, importance de la confidentialité dans l'utilisation de dossiers administratifs en vue d'établir des estimations pour domaine, et importance de la compatibilité des définitions lorsqu'on tente de combiner des données de diverses sources.

Les auteurs font ressortir très bien l'importance de prendre en considération l'estimation pour domaine à l'étape de la conception du plan de sondage, ce que font rarement d'autres auteurs qui s'intéressent particulièrement à l'estimation pour petite région. Comme le soulignent Singh, Gambino et Mantel, le statisticien pourra souvent construire des estimateurs pour domaine qui sont directs et convergents selon le plan s'il exécute avec soin l'étape de la conception. Je suis sûr que les concepteurs de plan de sondage tiennent compte de l'importance de la formation de grappes lorsqu'ils élaborent des enquêtes qui doivent servir à l'estimation pour domaine, mais il est réjouissant de voir qu'on en fait mention explicitement.

Les auteurs décrivent plusieurs types d'estimateur pour domaine. La classification de ces estimateurs fait ressortir le nombre de possibilités qui s'offrent au praticien. Les erreurs quadratiques moyennes théoriques peuvent nous renseigner sur les avantages relatifs des estimateurs. Afin d'illustrer ce genre de comparaisons, supposons un échantillon aléatoire simple de taille n prélevé dans une population qui est divisée en k domaines. Supposons qu'on connaît la taille des domaines et la valeur moyenne d'une variable auxiliaire, X , pour le domaine. Considérons les trois estimateurs par régression de la moyenne de domaine

$$\hat{\mu}^{(3)yt} = \bar{y}_{..} + (\mu_{xt} - \bar{x}_{..})b, \quad \text{et}$$

$$\hat{\mu}^{(2)yt} = \bar{y}_{.t} + (\mu_{xt} - \bar{x}_{.t})b,$$

$$\hat{\mu}^{(1)yt} = \bar{y}_{.t} + (\mu_{xt} - \bar{x}_{.t})b_t,$$

où

$$(\bar{x}_{..}, \bar{y}_{..}) = \sum_k N_{.t}^{-1} N_t(x_{.t}, y_{.t}),$$

$$(x_{.t}, y_{.t}) = \sum_{n_t}^{-1} N_t(x_{ij}, y_{ij}),$$

$$b_t = \left[\sum_{n_t}^{-1} (X_{ij} - \bar{x}_{.t})^2 \right]^{-1} \sum_{n_t}^{-1} (X_{ij} - \bar{x}_{.t})(Y_{ij} - \bar{y}_{.t}),$$

$$\times \sum_{n_t}^{-1} (X_{ij} - \bar{x}_{.t})(Y_{ij} - \bar{y}_{.t}),$$

$$b. = \left[\sum_k^{-1} N_{.t}^{-1} N_t n_t^{-1} \sum_{n_t}^{-1} (X_{ij} - \bar{x}_{.t})^2 \right]^{-1}$$

$$\times \sum_k^{-1} N_{.t}^{-1} N_t n_t^{-1} \sum_{n_t}^{-1} (X_{ij} - \bar{x}_{.t})(Y_{ij} - \bar{y}_{.t}),$$

n_t est le nombre d'observations dans le domaine t , N_t est le nombre d'éléments de la population contenus dans le domaine t , μ_{xt} est la moyenne de population de X pour le domaine t et $\mu_{x.}$ est la moyenne de X pour l'ensemble de la population. Selon la terminologie utilisée par les auteurs, le premier estimateur est un estimateur par régression direct, le deuxième, un estimateur direct modifié et le troisième, un estimateur synthétique. Nous avons

$$\text{EQM}\{\hat{\mu}^{(1)yt} | n_t\} = n_t^{-1} (1 + n_t^{-1}) V\{Y_{ij} - \beta_{\ell} X_{ij} | \ell = t\}$$

$$+ O(n_t^{-2}),$$

$$\text{EQM}\{\hat{\mu}^{(2)yt} | n_t\} = n_t^{-1} (1 + n_t^{-1}) V\{Y_{ij} - \beta X_{ij} | \ell = t\}$$

$$+ O(n_t^{-2}),$$

¹ W.A. Fuller, Distinguished Professor, Statistical Laboratory and Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa.

- SINGH, A.C., MANTTEL, H.J., et THOMAS, B.W. (1994). MPLSE à données chronologiques pour petites régions évaluées d'après des données d'enquête. *Techniques d'enquête*, 20, 35-46.
- SINGH, M.P., DREW, J.D., GAMBINO, J.G., et MAYDA, F. (1990). *Méthodologie de l'enquête sur la population active du Canada*. N° 71-526 au catalogue, Statistique Canada.
- SINGH, M.P., GAMBINO, J.G., et MANTTEL, H. (1992). Issues and options in the provision of small area statistics. *Small Area Statistics and Survey Designs*, (Vol. 1), (Eds. G. Kalton, J. Kordos et R. Platek), Warsaw: Central Statistical Office, 37-75.
- SINGH, M.P., et TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.
- VERMA, R.B.P., et BASAVARAJAPPA, J.G. (1987). Recent developments in the regression method for estimation of population for small areas in Canada. *Small Area Statistics*, (Eds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: Wiley and Sons, 46-61.
- U.S. STATISTICAL POLICY OFFICE (1993). Indirect Estimators in Federal Programs. Statistical Policy Working Paper 21. Préparé par le sous-comité sur Small Area Estimation, Federal Committee on Statistical Methodology.
- SWAIN, L., DREW, J.D., LAFRANCE, B., et LANCE, K. (1992). La création d'un registre des adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991. *Techniques d'enquête*, 18, 139-155.

BIBLIOGRAPHIE

- BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.
- BINDER, D., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, (Éds. P.R. Krishnaiah et C.R. Rao). New York: Elsevier Science, 187-211.
- BRACKSTONE, G.J. (1987a). Small area data: Policy issues and technical challenges. *Small Area Statistics*, (Éds. R. Platek, J.N.K. Rao, C.-E. Särndal et M.P. Singh), New York: Wiley and Sons, 3-20.
- BRACKSTONE, G.J. (1987b). Utilisations statistiques des données administratives: questions et défis. *Recueil: Symposium sur les utilisations statistiques des données administratives*, (Éds. J.W. Coombs et M.P. Singh), Statistique Canada, 5-17.
- CHOUDHRY, G.H., et RAO, J.N.K. (1989). Estimation de données régionales à l'aide de modèles qui combinent des séries chronologiques et des données transversales. *Recueil: Symposium sur l'analyse des données dans le temps*, (Éds. A.C. Singh et P. Whitridge), Statistique Canada, 71-80.
- COOMBS, J.W., et SINGH, M.P. (Éds.) (1987). *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada.
- DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labor Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 545-550.
- FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- FELLEGI, I.P. (1987). Allocation d'ouverture. *Recueil: Symposium sur les utilisations statistiques des données administratives*, (Éds. J.W. Coombs et M.P. Singh), Statistique Canada, 1-2.
- GHOSH, M., et RAO, J.N.K. (1993). Small area estimation: An appraisal. *A paraitre dans Statistical Science*.
- GHANGURDE, P.D., et SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 53-61.
- GONZALEZ, M.E. (1973). Use and evaluation of synthetic estimators. *Proceedings of the Social Statistics Section, American Statistical Association*, 33-36.
- GONZALEZ, M.E., et WAKSBERG, J. (1973). Estimation of the error of synthetic estimates. Présenté à la première réunion de l'Association Internationale des Statisticiens d'Enquête, Vienna, Austria.
- JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural Station Research Bulletin*, 304, 54-59.
- NORRIS, D., et PATON, D. (1991). L'enquête sociale générale canadienne: bilan des cinq premières années. *Techniques d'enquête*, 17, 245-260.
- OLKIN, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Série B*, 12, 241-255.
- PFEFFERMANN, D., et BURCK, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.
- PLATEK, R., et SINGH, M.P. (1986). Small Area Statistics, An International Symposium' 85 (articles contribues), Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, University of Ottawa, Canada.
- PRASAD, N.G.N., et RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or Domains). *Revue Internationale de Statistique*, 48, 3-18.
- RAO, J.N.K. (1986). Synthetic estimates, SPREE and best model based predictors. *Proceedings of the Conference on Survey Research Methodology in Agriculture*, American Statistical Association and National Agricultural Statistics Service, USDA, 1-6.
- ROYAL, R.M. (1979). Prediction models in small area estimation. Dans *Synthetic Estimates for Small Area, NIDA Research Monograph Series 24*, U.S. Department of Health, Education and Welfare.
- SÄRNDAL, C.-E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.
- SÄRNDAL, C.-E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHAIKLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Area, NIDA Research Monograph Series 24*, U.S. Department of Health, Education and Welfare. Library of Congress catalogue number 79-600067, 36-53.
- SCHAIKLE, W.L. (1992). Use of small area estimators in U.S. federal programs. *Small Area Statistics and Survey Designs*, (Éds. G. Kalton, J. Kordos et R. Platek). Warsaw: Central Statistical Office, 95-114.
- SCHMIDT, R.C. (1952). Short-cut methods for estimating county populations. *Journal of the American Statistical Association*, 47, 232-238.
- SINGH, A.C., et MANTTEL, H.J. (1991). Estimation composite par espace d'états pour les petites régions. *Recueil: Symposium 91, Questions spatiales liées aux statistiques*, Statistique Canada, 21-30.
- SINGH, Gambino et Mantel: Les petites régions: problèmes et solutions

la taille de l'échantillon. Si les échantillons des différentes périodes se chevauchent parfaitement, la mise en moyenne ne se traduit par aucun gain d'efficacité. Cependant, pour d'autres plans de renouvellement, ce procédé pourrait plus que doubler la taille de l'échantillon pour des estimations pour domaine. Mais les estimations groupées posent un problème théorique en ceci qu'elles représentent une valeur moyenne du paramètre étudié (par ex., niveau de chômage) pour une période de trois mois par exemple.

Dans l'estimation composite, on combine l'estimateur de plan de la période courante avec l'estimateur composite de la période précédente, ce dernier étant redressé au moyen d'une estimation de la variation basée sur l'échantillon commun. Ce procédé a été appliqué pour autre chose que l'estimation pour petites régions par Jessen (1942) et Patterson (1950), pour ne nommer que ceux-là. Binder et Hidiroglou (1988) font un examen de la question. Les poids nécessaires à la combinaison sont ordinairement des estimations des poids optimaux dans l'hypothèse que ceux-ci ne varient pas dans le temps. Ces poids tributaires des données ont pour inconvénient de produire des estimations incohérentes pour diverses caractéristiques et les sommes de ces caractéristiques.

Dans l'estimation pour petites régions, on a recours depuis peu à des méthodes basées sur des séries temporelles pour les enquêtes périodiques. On modélise la relation entre les paramètres étudiés pour différentes périodes, puis on se sert du modèle afin d'accroître l'efficacité des estimations pour le passage courant. Dans la plupart des cas, il faut aussi tenir compte, par la modélisation ou autrement, de la non-indépendance des échantillons des divers passages, condition attribuable au plan de renouvellement de l'échantillon. Cette approche fondée sur les séries chronologiques est traitée dans les ouvrages suivants: Choudhry et Rao (1989); Pfeffermann et Burck (1990); Singh, Mantel et Thomas (1994) et Singh et Mantel (1991). Tous ces ouvrages présentent des versions généralisées du modèle de Fay-Herriot selon lesquelles les paramètres de régression, les effets de petite région et les erreurs d'enquête varient selon divers modèles chronologiques. Le vecteur d'estimations régionales calculé par cette méthode peut être exprimé comme la moyenne pondérée du vecteur des estimations de plan et du vecteur des estimations synthétiques, celles-ci étant fondées sur des données de périodes antérieures et les valeurs courantes des covariables; cependant, la matrice des poids sera très rarement une matrice diagonale, de sorte que l'estimateur pour une petite région quelconque dépendra généralement aussi bien des estimations de plan et des estimations synthétiques d'autres petites régions.

7. CONCLUSION

Les concepteurs de plan de sondage doivent surmonter beaucoup d'obstacles pour produire des estimations pour domaine qui soient précises et actuelles. Ils doivent tout d'abord convaincre les promoteurs d'enquêtes et les chargés de programme qu'un système conçu spécialement pour produire des estimations nationales et infra-nationales ne peut répondre à tous les besoins en données régionales. On

peut réaliser des gains d'efficacité appréciables (selon les enquêtes) au niveau du domaine avec très peu de perte aux niveaux supérieurs. Nous avons souligné la nécessité d'élaborer une approche globale, par laquelle on pourrait définir un degré de fiabilité pour les domaines planifiés comme pour les régions de niveau supérieur par des réparations intermédiaires et par laquelle on pourrait réduire la formation de grappes dans le but d'améliorer les estimations relatives aux domaines non planifiés. Il convient de noter que bon nombre de domaines dits planifiés à l'étape de la conception du plan de sondage peuvent se transformer en domaines non planifiés (révisés) au fil du temps lorsqu'il s'agit d'enquêtes permanentes.

L'approche globale devrait aussi permettre de considérer les estimateurs de plan pour de grands domaines aussi bien que les estimateurs de modèle pour de petits domaines. L'estimateur de modèle devrait être préféré à l'estimateur de plan seulement si l'erreur quadratique moyenne du premier (variance de plan + biais²) peut être estimée et qu'elle est inférieure à la variance correspondante de l'estimateur de plan **par une marge suffisante**. Il devrait y avoir des estimations de l'erreur quadratique moyenne pour chacun des domaines **pris individuellement**. Les organismes statistiques peuvent choisir de grouper des domaines semblables ou de grouper des estimations de périodes différentes qui se rapportent au même domaine. Ils peuvent même supprimer les estimations touchant certains domaines à cause de leur faible degré de fiabilité ou pour des raisons de protection du secret statistique.

La deuxième tâche qui attend les statisticiens est d'expliquer aux utilisateurs les divers types de mesures de fiabilité qui peuvent s'appliquer à différentes séries d'estimations provenant de la même enquête. Espérons qu'avec la poursuite des recherches sur la validation de modèle et de meilleurs estimations de l'erreur quadratique moyenne, les concepteurs hésiteront moins à utiliser des estimateurs de modèle pour les petits domaines. D'ici là, on devra se servir des estimateurs de modèle avec circonspection même si leurs coefficients de variation sont beaucoup plus petits. Les recensements, avec les fichiers administratifs en appui, demeureront vraisemblablement la principale source de données socio-économiques pour les petites régions, surtout dans les pays où il se fait un recensement de la population et du logement à tous les cinq ans. En outre, on continuera de se préoccuper des problèmes conceptuels que posent les données des fichiers administratifs tant que les organismes statistiques n'auront pas la chance de participer à l'élaboration des formules qui servent à recueillir ces données. En attendant, on ne peut exploiter pleinement cette source de données extrêmement riche que constituent les fichiers administratifs; ce qui est dommage pour l'analyse statistique et surtout pour l'estimation pour domaine.

REMERCIEMENTS

Nous sommes reconnaissants à Jon Rao d'avoir agi comme rédacteur associé pour cet article et nous remercions aussi les rapporteurs pour leurs nombreux conseils utiles.

e_a sont indépendants et non corrélés pour a , σ^2 est inconnu et les v_a^2 sont supposés connus (en pratique, ils devraient être estimés). Pour une valeur donnée de σ^2 , on peut calculer les poids optimaux nécessaires à la combinaison de $X_a^{des,a}$ et de X_a^{β} . On calcule une estimation de σ^2 à l'aide de la méthode des constantes d'ajustement, puis on introduit cette estimation dans la formule des poids optimaux. Pour parer à la défaillance du modèle, on prévoit une troncation de l'estimation calculée si l'écart entre celle-ci et l'estimation directe est supérieur à un multiple déterminé de v_a . Schauble (1979) et Battese et Fuller (1981) considèrent aussi des poids optimaux λ_a (en 6.12) estimés empiriquement selon des modèles à effets aléatoires semblables pour les totaux de petites régions. Prasad et Rao (1990) définissent un estimateur de l'erreur quadratique moyenne de l'estimateur de Fay-Herriot qui tient compte de l'estimation des composantes de la variance. Kott (1989) propose un estimateur convergent selon le plan de l'erreur quadratique moyenne mais constate que cet estimateur est très instable.

Une autre méthode consiste à calculer les poids au moyen de données historiques; l'avantage de cette solution est que les poids calculés de cette manière peuvent être plus stables que des poids estimés à l'aide des données d'enquête courantes; néanmoins, cela suppose l'hypothèse que les poids optimaux sont stables.

Remarque: Dans l'estimation tribunaire de la taille de l'échantillon, les poids peuvent dépendre de la taille observée du sous-échantillon s_a mais non des valeurs de la variable étudiée. Cette absence de relation de dépendance entre les poids et la variable étudiée présente des avantages et des inconvénients. Un des avantages est que les mêmes poids servent à l'estimation de totaux pour toutes les variables étudiées. On n'a à les calculer qu'une seule fois. Un avantage plus important encore est que l'estimation de la somme de deux variables est égale à la somme des estimations de chaque variable. Du côté des inconvénients, les poids ne tiennent pas compte directement de la précision de l'estimateur de plan pour la variable étudiée ou de la grande probabilité du biais de l'estimateur synthétique.

Combinaison de données de diverses périodes: Dans les enquêtes à passages répétées, il est courant de grouper les données de plusieurs passages dans le but d'accroître la précision des estimations. Selon le plan de renouvellement utilisé dans ces enquêtes, on peut réussir à accroître sensiblement la fiabilité des estimations. Cette technique de groupement ou de mise en moyenne sur plusieurs périodes suscite donc un intérêt particulier en ce qui concerne l'estimation pour domaine, caractérisée par un degré de précision modeste. En ce qui regarde l'estimation pour domaine dans l'enquête sur la population active du Canada, il est pratiqué couramment d'utiliser un estimateur tribunaire de la taille de l'échantillon basé sur une moyenne trimestrielle des estimations du nombre de personnes occupées et du nombre de personnes en chômage. Grâce au plan de renouvellement de six mois utilisé dans cette enquête, la mise en moyenne sur trois mois permet d'accroître du tiers

Des estimateurs de ce genre ont été proposés antérieurement. Drew, Singh et Choudhry (1982) ont proposé, par exemple, l'estimateur suivant (tribunaire de la taille de l'échantillon):

$$Y^{ssd,r,a} = \lambda_a Y^{r,a} + (1 - \lambda_a) Y^{syn,r,a}, \quad (6.11a)$$

$$\lambda_a = \begin{cases} 1 & \text{si } N_{e,a} \geq \delta N_a \\ N_{e,a} / \delta N_a & \text{dans le cas contraire} \end{cases} \quad (6.11b)$$

et la valeur δ est choisie subjectivement pour régler la contribution de la composante synthétique. Särndal (1984) a proposé l'estimateur suivant

$$Y^{ssd,reg,a} = \lambda_a Y^{sreg,a} + (1 - \lambda_a) Y^{syn,reg,a}, \quad (6.12)$$

où $\lambda_a = N_{e,a} / N_a$. Rao (1986) a proposé de modifier cet estimateur dans le sens suivant: λ_a aurait la valeur 1 si $N_{e,a} \geq N_a$. Särndal et Hidiroglou (1989) ont amélioré la proposition de Rao en ajoutant la relation suivante: $\lambda_a = (N_{e,a} / N_a)^{h-1}$ si $N_{e,a} < N_a$, la valeur h étant choisie de façon rationnelle pour régler la contribution de la composante synthétique.

Dans la pratique, ce qui compte le plus lorsqu'on utilise des estimateurs de ce genre est le biais de la composante synthétique. Cette composante devrait être affectée d'un poids tel que le biais se situe dans des limites raisonnables. Par exemple, on se sert actuellement de l'estimateur tribunaire de la taille d'échantillon de Drew, Singh et Choudhry (1982) dans l'enquête sur la population active du Canada pour produire des estimations pour domaine, compte tenu de ce que l'estimation par régression généralisée remplace l'estimation par quotient et que $\delta = 2/3$. Pour une majorité de domaines, le poids attribué à la composante synthétique est nul puisque l'estimateur direct satisfait à lui seul les exigences de fiabilité. Pour les autres domaines, le poids de la composante synthétique est d'environ 10% en moyenne et ne dépasse jamais 20%. Selon le biais qu'on est prêt à accepter, la valeur de δ peut se situer dans l'intervalle $[2/3, 3/2]$ pour la plupart des cas.

La troisième méthode de pondération met les poids dans un rapport de dépendance avec les données. Les poids optimaux nécessaires à la combinaison de deux estimateurs dépendent généralement de l'erreur quadratique moyenne des estimateurs et de leur covariance. Ces valeurs sont le plus souvent inconnues mais elles peuvent être estimées à partir des données. Dans le cas des estimateurs combinés, il faudrait normalement modéliser le biais de la composante synthétique. On trouve dans Fay et Herriot (1979) un exemple très connu d'application de cette méthode, qui est un des premiers du genre dans la littérature statistique. Fay et Herriot modélisent les biais des estimateurs synthétiques pour les petites régions comme des effets aléatoires indépendants ayant une variance inconnue mais fixe. En termes plus précis, si $Y_a^{des,a}$ est l'estimateur de plan, ils considèrent alors le modèle $Y_a = X_a^{\beta} + \alpha_a$ et l'estimateur $Y_a^{des,a} = Y_a + \epsilon_a$, où $\alpha_a \sim (0, \sigma^2)$, $\epsilon_a \sim (0, v_a^2)$, et α_a et

Une forme plus courante d'estimateur synthétique est celle qui repose sur la stratification ou la stratification a posteriori, soit

$$Y_{syn,sl,m,a} = \sum_h N_{h,a} \sum_{i \in s_h} w_i y_i / \sum_h w_i = \sum_h N_{h,a} \bar{y}_h.$$

Comme dans le cas des estimateurs directs, les estimateurs synthétiques peuvent utiliser des données auxiliaires en plus des effectifs N_a et $N_{h,a}$. Par exemple, les estimateurs synthétiques par quotient basés sur une covariable x sont définis le plus souvent sous la forme

$$Y_{syn,a} = X_a Y_e / X_e \quad \text{et} \quad Y_{syn,sl,a} = \sum_h X_{h,a} Y_{h,e} / X_{h,e}, \quad (6.6)$$

où $Y_e = \sum_{i \in s} w_i y_i$ est l'estimateur par facteur d'extension du total de population pour y et $Y_{h,e} = \sum_{i \in s_h} w_i y_i$. X_e et $X_{h,e}$ sont définis de la même manière. Ces estimateurs ont été étudiés par Gonzalez (1973), Gonzalez et Waksburg (1973) et Changuarde et Singh (1977, 1978), pour n'en nommer que quelques-uns.

Singh et Tessier (1976) ont proposé une autre forme d'estimateur synthétique par quotient, où X remplace X_e :

$$Y_{syn,r,a} = X_a Y_e / X. \quad (6.7)$$

Les deux estimateurs, $Y_{syn,r,a}$ et $Y_{syn,a}$, ont le même biais synthétique et le biais de $Y_{syn,r,a}$ est négligeable pour de grands échantillons. Le choix entre l'un et l'autre de ces estimateurs dépend de ρ , le coefficient de corrélation de Y_e et X_e . On peut montrer que pour de grands échantillons, $V(Y_{syn,r,a}) \leq V(Y_{syn,a})$ si $\rho \geq 0.5 c_x / c_y$, où c_x et c_y sont les coefficients de variation de X_e et de Y_e respectivement. Dans la plupart des cas, lorsque ρ est élevé ou que la population est asymétrique, on préférera $Y_{syn,r,a}$; cependant, si c_x est élevé et que la corrélation n'est pas trop forte, $Y_{syn,r,a}$ pourra être un meilleur choix.

Dans certains cas, il peut exister de l'information sur une autre variable auxiliaire (2) en plus de x . On peut alors construire un estimateur synthétique par quotient à deux variables:

$$Y_{(2)_{syn,r,a}} = \gamma_a X_a Y_e / X_e + (1 - \gamma_a) Z_a Y_e / Z_e, \quad (6.8)$$

où la valeur de γ_a est choisie judicieusement. On peut aussi envisager un estimateur synthétique par quotient à plusieurs variables si l'on se fonde sur l'article d'Olkin (1958).

L'estimation synthétique par régression ressemble à l'estimation synthétique par quotient,

$$Y_{syn,reg,a} = \beta X_a,$$

$$\beta = \frac{\sum_{i \in s} v_i^{-1} w_i y_i x_i}{\sum_{i \in s} v_i^{-1} w_i x_i x_i} \quad (6.9)$$

Comme dans le cas des estimateurs directs, l'estimateur synthétique par régression peut être appliqué à des strates initiales ou à des strates formées a posteriori. Royall (1979) a proposé un variante de cet estimateur, soit $Y_{syn,Roy,a} = \sum_{i \in s_a} y_i + \beta(X_a - \sum_{i \in s_a} x_i)$, où la somme des valeurs y qui se rapportent uniquement aux unités non incluses dans l'échantillon est estimée de façon synthétique.

Remarque: Les estimateurs directs modifiés présentés dans la section 6.1 peuvent aussi être exprimés comme des estimateurs synthétiques par quotient ou des estimateurs synthétiques par régression comportant un facteur de compensation du biais basé sur le plan. Nous pouvons écrire par exemple l'expression $Y_{sreg,a} = Y_{syn,reg,a} + (Y_a - \beta X_a)$, où $Y_a - \beta X_a$ est la valeur estimée du biais de $Y_{syn,reg,a}$. De même, nous pouvons exprimer $Y_{sreg,a}$ comme l'estimateur de Royall, $Y_{syn,Roy,a}$, accompagné d'un facteur de compensation du biais basé sur le plan.

Purcell et Kish (1980) étudient une autre forme d'estimation synthétique qu'ils appellent SPREB (pour "structure preserving estimation") et qui s'applique à l'estimation d'effectifs pour petites régions. Selon cette méthode, on combine des données antérieures détaillées, provenant peut-être d'un recensement, avec des estimations d'enquête courantes moins détaillées dans le but d'obtenir des estimations détaillées de valeurs courantes. Cette méthode suppose que certains rapports parmi les données détaillées ne varient pas dans le temps.

Estimateurs combinés: L'estimateur combiné est défini comme la moyenne pondérée d'un estimateur de plan et d'un estimateur synthétique, c'est-à-dire

$$Y_{com,a} = \lambda_a Y_{des,a} + (1 - \lambda_a) Y_{syn,a}, \quad (6.10)$$

où la valeur de λ_a est choisie judicieusement. L'estimateur combiné vise à faire l'équilibre entre la possibilité de l'estimateur synthétique d'être biaisé et l'instabilité de l'estimateur de plan. Il existe trois grandes méthodes pour déterminer les poids λ_a en (6.10): ces poids peuvent être établis à l'avance, ils peuvent dépendre de la taille de l'échantillon ou encore, ils peuvent dépendre des données. La première méthode de pondération, et la plus simple, consiste à fixer des poids à l'avance (par exemple, utiliser une moyenne simple). Or, cette méthode ne tient aucunement compte de la précision réelle de l'estimateur de plan. L'estimateur de plan pour la petite région a est plus précis pour certains échantillons réels que pour d'autres. Le poids attribué à l'estimateur de plan devrait refléter cette différence. La deuxième grande méthode de pondération des composantes de l'estimateur combiné a un rapport de dépendance avec la taille de l'échantillon; en effet, les poids sont fonction du rapport $N_{e,a}/N_a$. Une variante de cette méthode, que nous n'examinons pas ici, a plutôt un rapport de dépendance avec les valeurs empiriques d'une covariable x ; par exemple, les poids peuvent être fonction de $X_{des,a}/X_a$ ou de $S_{x,a}^2/\sigma_x^2$, où $S_{x,a}^2$ est la variance empirique de $X_{des,a}$, étant donné $N_{e,a}$ ou une autre caractéristique pertinente de l'échantillon réel, et σ_x^2 est la variance inconditionnelle de $X_{des,a}$.

6.1 Estimateurs de plan

Estimateurs directs: Les estimateurs directs pour petite région reposent sur des données d'enquête qui proviennent uniquement de la petite région; à ces données peuvent s'ajouter parfois des données auxiliaires tirées du recensement ou de fichiers administratifs. La forme la plus simple de l'estimateur direct d'un total est l'estimateur par facteur d'extension,

$$Y_{e,a} = \sum_{i \in s_a} w_i y_i \quad (6.1)$$

où s_a est la portion de l'échantillon contenue dans la petite région a et w_i est le poids de sondage de l'unité i . Cet estimateur est non biaisé; cependant, il peut être caractérisé par une forte variabilité du fait que la taille d'échantillon dans la région a est aléatoire.

Si la taille de la population, N_a , est connue, on peut se servir d'un estimateur de stratification a posteriori,

$$Y_{pst,a} = N_a \sum_{i \in s_a} w_i y_i / \sum_{i \in s_a} w_i = N_a Y_{e,a} / N_{e,a} = N_a \bar{y}_{e,a}. \quad (6.2)$$

Cet estimateur est plus stable que l'estimateur par facteur d'extension; cependant, il peut être entaché d'un biais d'estimation par quotient dans les enquêtes complexes.

Si l'échantillonnage est stratifié et que $N_{h,a}$, l'effectif de la population contenu dans la strate h et la petite région a , est connu, on peut utiliser un autre estimateur de stratification a posteriori: $Y_{st,pst,a} = \sum_h (N_{h,a} \sum_{i \in s_{h,a}} w_i y_i / \sum_{i \in s_{h,a}} w_i) = \sum_h N_{h,a} Y_{h,e,a} / N_{h,e,a}$. Il peut aussi s'agir de strates formées a posteriori au lieu de strates initiales. L'estimation par quotient ressemble à l'estimation de stratification a posteriori avec cette différence qu'on utilise une variable auxiliaire au lieu des effectifs N_a et $N_{h,a}$. Par exemple, si x est une covariable pour laquelle on connaît le total de petite région, X_a , ou le total de strate de petite région, $X_{h,a}$, nous pouvons définir les estimateurs par quotient

$$Y_{r,a} = X_a R_a \quad \text{et} \quad Y_{st,r,a} = \sum_h X_{h,a} R_{h,a}, \quad (6.3)$$

où $R_a = Y_{e,a} / X_{e,a}$ est une estimation du rapport Y_a / X_a et $R_{h,a} = Y_{h,e,a} / X_{h,e,a}$.

Par ailleurs, l'estimateur par régression tente d'expliquer la différence entre la valeur de covariables pour l'effectif d'une petite région et celle pour un sous-échantillon au moyen d'une équation de régression estimée qui met en relation la variable étudiée y et les covariables x . Un des avantages de l'estimation par régression est qu'elle s'étend facilement à des covariables vectorielles. L'estimateur est défini

$$Y_{reg,a} = Y_a + \beta_a (X_a - X_a), \quad (6.4)$$

où Y_a peut être un estimateur par facteur d'extension ou un estimateur de stratification a posteriori, X_a doit être calculé de la même manière que Y_a , et $\beta_a = \sum_{i \in s_a} w_i y_i x_i' / \sum_{i \in s_a} w_i^2 x_i x_i'$, où w_i désigne les poids donnés pour la régression. Notons que $\beta_a = R_a$ lorsque x est une grandeur scalaire et $v_i = x_i$. Lorsque Y_a et X_a sont des estimateurs par facteur d'extension, l'estimateur (6.4) est aussi appelé estimateur par régression généralisé. Il est approximativement non biaisé selon le plan à la condition qu'il en soit de même pour Y_a et X_a .

Comme l'estimateur par quotient, l'estimateur par régression peut être appliqué à des strates initiales ou à des strates formées a posteriori.

Estimateurs directs modifiés: Les estimateurs directs modifiés peuvent utiliser des données d'enquête qui ne proviennent pas du domaine étudié; cependant, ils demeurent des estimateurs approximativement non biaisés selon le plan. Selon notre définition, l'estimateur direct modifié est un estimateur direct qui comporte un facteur de correction synthétique pour biais de modèle; comme l'espérance du facteur de correction serait approximativement nulle par rapport au plan, l'estimateur modifié est approximativement non biaisé selon le plan si l'estimateur direct est aussi approximativement non biaisé selon le plan. On a un exemple d'estimateur direct modifié si on remplace β_a dans (6.4) par un estimateur synthétique, $\beta_a = \sum_{i \in s} v_i^{-1} w_i y_i x_i' / \sum_{i \in s} v_i^{-1} w_i x_i x_i'$, nous désignons cet estimateur par $Y_{reg,a}^{sr}$. β sera généralement plus stable que β_a ; le choix entre l'un et l'autre dépendra du rapport entre la grandeur de la variance de β_a et la variance des valeurs β_a pour les régions a . Une solution intermédiaire serait d'opter pour une moyenne pondérée, $\lambda_a \beta_a + (1 - \lambda_a) \beta$, où la valeur de λ_a est choisie judicieusement; le choix d'une valeur de λ_a est examiné dans la partie de la section 6.2 qui porte sur les estimateurs combinés. On a un autre exemple d'estimateur direct modifié si on remplace β_a dans (6.4) par β_a dans (6.4) par $R = Y_e / X_e$; notons que R est un cas particulier de β lorsque x est une grandeur scalaire et $v_i = x_i$.

6.2 Estimateurs indirects

Estimateurs synthétiques: Les méthodes d'estimation synthétique reposent sur l'hypothèse selon laquelle la petite région ressemble d'une certaine manière à une autre région, souvent plus grande, dans laquelle elle est contenue. Les estimations relatives à cette autre région seraient généralement plus précises que les estimations relatives à la petite région. L'estimateur synthétique serait alors un estimateur à faible variance, qui pourrait toutefois être fortement biaisé si l'hypothèse mentionnée plus haut ne se vérifiait pas. L'une des formes les plus simples d'estimateur synthétique repose sur l'hypothèse que la moyenne pour petite région est égale à la moyenne globale. On a ainsi l'estimateur synthétique de moyenne

$$Y_{syn,m,a} = N_a \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i = N_a \bar{y}. \quad (6.5)$$

tout le moins en théorie, favoriser la production d'un grand nombre d'estimations régionales. Si nous avons un échantillon mensuel de 59,000 ménages et si nous supposons qu'il faut, par exemple, 100 ménages par mois pour produire des estimations trimestrielles fiables, on peut diviser le pays en quelque 600 régions non chevauchantes, chacune contenant avec certitude un nombre suffisant d'unités de l'échantillon. De la même manière, en groupant des régions, on obtiendra des territoires qui contiendront un nombre suffisant d'unités pour produire des estimations mensuelles fiables.

Il existe diverses méthodes de répartition de l'échantillon. Dans une perspective du haut vers le bas, on détermine tout d'abord la taille de l'échantillon provincial; on répartit ensuite l'échantillon entre les régions infraprovinciales. Il se peut toutefois que la taille de l'échantillon provincial empêche que le degré de fiabilité requis pour les estimations infraprovinciales soit respecté. Dans une perspective du bas vers le haut, on répartit tout d'abord l'échantillon entre les régions infraprovinciales de manière à répondre aux objectifs de fiabilité pour chaque région. En conséquence, ces régions devraient normalement avoir des tailles d'échantillon comparables. Avec cette méthode, il se peut que la taille de l'échantillon provincial soit plus grande que celle déterminée par la méthode précédente. Quelle que soit la méthode utilisée, il faudra normalement apporter des corrections aux chiffres initiaux. La répartition finale sera une sorte de compromis entre la répartition proportionnelle et la répartition uniforme. En pratique, le concepteur de plan de sondage doit se livrer à un exercice complexe où il doit tenir compte à la fois des conditions de fiabilité au niveau provincial, des exigences au niveau infraprovincial pour un ou plusieurs groupes de régions, du coût total de l'enquête et des conditions sur le terrain. La méthode de répartition qui a été choisie pour la version remaniée de l'EPA peut être utile pour d'autres enquêtes. La répartition de l'échantillon s'est faite en deux étapes: on a tout d'abord réparti un échantillon de base de 42,000 ménages de manière à obtenir de bonnes estimations aux niveaux national et provincial; on a ensuite réparti le reste de l'échantillon de manière à obtenir des estimations infraprovinciales le plus justes possible. Cette formule de répartition permettra d'obtenir des estimations fiables pour presque tous les domaines planifiés. Cette solution de compromis s'est traduite par des pertes minimes au niveau provincial. Par exemple, le c.v. théorique pour les "personnes en chômage" en Ontario et au Québec est 3,2 et 3,0% respectivement, au lieu de 2,8 et 2,6%. Les pourcentages correspondants pour le Canada sont 1,51 et 1,36. Si on vise l'optimisation au niveau provincial, on obtient des c.v. pouvant atteindre 17,7% pour les régions de l'A.-C. Grâce à la formule de répartition décrite ci-dessus, ces mêmes c.v. ne dépassent pas 9,4%.

Redistribution de l'échantillon: Le concepteur jouit normalement d'une certaine marge de manœuvre pour déplacer des unités de l'échantillon d'une région à une autre. Par exemple, si on réduit de 1,000 ménages la taille de l'échantillon d'une grande province et qu'on augmente d'autant la taille de l'échantillon d'une petite province, la

qualité des estimations pour la grande province s'en trouvera peu affectée alors que celle des estimations pour la petite province en sera grandement améliorée. On peut effectuer des transferts semblables à l'intérieur d'une province.

5.5 Autres considérations

Modification de la définition des petites régions: Les concepteurs de plan de sondage doivent se faire à l'idée que les définitions de domaines planifiés peuvent changer au cours de l'existence d'une enquête et qu'ils peuvent devoir considérer de nouveaux domaines comme des domaines non planifiés. Par exemple, il se peut très bien que la définition des régions de l'Assurance-chômage soit modifiée dans les deux ou trois années qui vont suivre la mise en application du nouveau plan de sondage de l'EPA en 1995. La meilleure chose que le concepteur pourra faire pour parer à cette difficulté à l'étape de l'élaboration sera de choisir comme unités de base des régions dites normales (par ex., des régions propres au recensement dont la définition varie peu) et d'espérer que les régions redéfinies correspondent à un groupement de régions normales. C'est la solution qui est préconisée dans la version remaniée de l'EPA.

Une autre solution consisterait à adopter une méthode de mise à jour. Cela nécessiterait le tirage de nouvelles unités; ce tirage serait effectué de manière à maximiser le chevauchement des unités tirées initialement et des nouvelles unités. De cette manière, on réduirait au maximum le nombre de nouvelles unités à inscrire sur la liste. On réduirait aussi au maximum les autres contraintes opérationnelles, comme la nécessité d'emboucher d'autres intervieweurs.

6. ESTIMATION

Cette section a pour but d'exposer quelques-unes des méthodes d'estimation de totaux pour petites régions. Nous ne cherchons pas ici à faire une analyse exhaustive de ces méthodes; nous décrivons plutôt l'évolution des recherches sur l'estimation pour petites régions. Pour une analyse détaillée de la question, le lecteur est prié de consulter le tout récent article de Ghosh et Rao (1993). Pour faciliter notre exposé, nous allons classer les méthodes d'estimation pour petites régions en deux groupes. (Il existe beaucoup d'autres possibilités de classement). Le premier groupe comprend les estimateurs de plan, c.-à-d. des estimateurs (approximativement) non biaisés selon le plan; dans ce groupe on trouve les estimateurs directs et les estimateurs directs modifiés. Comme nous l'avons mentionné plus haut, les estimateurs de plan laissent souvent à désirer à cause de leur variance élevée, qui s'explique par le faible effectif des petites régions (ou même un effectif nul). Le second groupe est désigné comme le groupe des estimateurs indirects (ou estimateurs de modèle) et il comprend les estimateurs synthétiques et les estimateurs combinés. Certains de ces estimateurs font l'objet d'une comparaison empirique dans une version antérieure de cet article; voir Singh, Gambino et Mantel (1992).

mise à jour du registre des adresses post-censitaire en vue des enquêtes et des recensements futurs.

En ce qui concerne la deuxième solution, on a envisagé la modification d'unités et la réduction du nombre de degrés d'échantillonnage pour les régions non urbaines en gardant en mémoire l'idée qu'une réduction de la formation en grappes se traduit par de meilleures estimations régionales. Étant donné l'évolution qu'ont connue les méthodes de collecte de données (de l'interview directe à l'interview téléphonique à l'interview assistée par ordinateur), les analyses qui visaient auparavant à trouver un équilibre entre le coût d'une enquête et la variance désirée sont révolues. Plus de 80 pour cent des interviews de l'EPA sont maintenant effectuées par téléphone. Avec l'accroissement du nombre des interviews téléphoniques et la diminution connexe du nombre des déplacements, il est presque toujours possible maintenant de supprimer le premier degré d'échantillonnage et de prélever directement des SD.

5.3 Stratification

Une méthode de stratification conforme à l'esprit de l'analyse qui s'est faite précédemment sur la taille de l'UPB consiste à remplacer les grandes strates par de nombreuses petites strates dans l'espoir qu'un domaine redéfini ou qu'un domaine non planifié renfermerait des strates presque complètes. La taille des domaines sera ainsi plus stable.

Il peut y avoir plusieurs régions en chevauchement pour lesquelles on a besoin d'estimations. Par exemple, chaque province canadienne est divisée en régions économiques (RE) et en régions de l'Assurance-chômage (RAC). Une solution consiste à considérer tous les secteurs créés par l'intersection des régions comme des strates. Ainsi, au Canada, le recoupement des 71 RE et des 61 RAC produit en tout 133 régions en chevauchement, ce qui est raisonnable. Dans certains cas toutefois, le nombre de régions chevauchantes pourrait être trop élevé pour que les opérations soient efficaces. En outre, certains des secteurs pourraient avoir un effectif minime, ce qui en ferait des strates inutilisables.

En combinant la réduction de la formation en grappes et la diminution de la taille des strates, nous espérons avoir un plan de sondage qui pourra mieux répondre aux exigences des petites régions. Par exemple, les plans de ce genre devraient offrir toute la souplesse voulue pour répondre efficacement aux exigences des RE et des RAC et traiter les changements de définition des régions.

5.4 Répartition

Si nous connaissons d'avance les définitions des petites régions, nous pourrions peut-être considérer ces régions comme des domaines planifiés et en tenir compte au moment de la conception de l'enquête. Le concepteur peut s'appliquer à répartir l'échantillon dans les petites régions de manière à rendre possible la production d'estimations fiables. Pour de grandes enquêtes comme l'Enquête sur la population active du Canada, cette approche peut, à

5.2 Degrés d'échantillonnage et unités d'échantillonnage

Chaque des dix provinces du Canada est divisée en régions économiques (RE), lesquelles, en vertu de l'EPA, sont subdivisées en régions auto-représentatives (grandes villes et villes de taille moyenne) et en régions non auto-représentatives (le reste de la RE). La stratification et l'échantillonnage s'effectuent à l'intérieur de ces régions, et le nombre de degrés d'échantillonnage de même que les unités d'échantillonnage varient selon le type de région. Par exemple, le nombre de degrés d'échantillonnage est de trois dans les régions non auto-représentatives mais seulement deux dans les régions auto-représentatives. Pour une description détaillée du plan d'échantillonnage de l'EPA, prière de se référer à Singh et coll. (1990).

Les bases aréolaires ont rapport généralement à l'échantillonnage en grappes, c'est-à-dire que les unités d'échantillonnage du premier degré sont habituellement des superflèches qui contiennent un certain nombre d'unités du deuxième degré. Si on dispose d'une liste des unités du deuxième degré, on peut alors effectuer un échantillonnage à même la liste, ce qui donne un échantillon comportant un moins grand nombre de grappes. Nous aurons ainsi non seulement des estimations de meilleure qualité (grâce à des effets de plan moindres), mais aussi de meilleures estimations régionales pour les domaines non planifiés. La seconde conséquence est vraie car une répartition plus uniforme de l'échantillon à de fortes chances de "placer" des unités échantillonnées dans un domaine spatial non planifié. Inversement, dans le cas d'un plan avec formation de grappes, il arrive souvent qu'un domaine renferme un nombre surflissant d'unités de l'échantillon du fait qu'il contient des grappes échantillonnées, tandis qu'un autre domaine semblable ne contiendra pas suffisamment de grappes pour produire de bonnes estimations.

Nous avons examiné deux solutions pour réduire la formation de grappes dans l'EPA: i) remplacer la base aréolaire (avec plan à deux degrés) pour les grandes villes par une nomenclature fondée sur le Registre des adresses ou ii) réduire le nombre de degrés d'échantillonnage pour les régions rurales et les petits centres urbains. Le Registre des adresses, qui a été créé dans le but d'améliorer le taux de couverture du recensement du Canada de 1991 (Swain, Drew, Lafrance et Lance 1992), contient des adresses et des numéros de téléphone ainsi que de l'information géographique se rapportant à des logements pour chaque secteur de dénombrement (SD). La première solution examinée consistait à tirer un échantillon aléatoire simple stratifié de logements dans le Registre des adresses. À cet échantillon pouvait s'ajouter un second tiré d'une base de croissance qui contient des logements qui ne figurent pas dans le registre des adresses post-censitaire. Seulement, la manière de traiter la croissance est devenue le principal obstacle dans l'application de la première solution car il n'a pas été possible d'élaborer et de tester une méthode d'un bon rapport coût-efficacité dans les délais requis par le projet de remaniement de l'EPA. Néanmoins, on travaille actuellement à l'élaboration d'une méthode de

besoins en données régionales en se fondant sur la demande antérieure. Quant aux enquêtes ponctuelles, les concepteurs devraient inclure la définition des besoins dans le processus d'établissement des objectifs de l'enquête. Par conséquent, dans l'un et l'autre cas, les concepteurs devraient fixer un degré de précision non seulement pour les estimations nationales ou provinciales, mais aussi pour les domaines étudiés.

La première étape dans la fourniture de données régionales dépendra dans quelle mesure les domaines auront été définis à l'avance, de sorte qu'ils puissent être considérés comme des domaines planifiés au moment de la conception (ou du remaniement) de l'enquête. Si, pour des considérations budgétaires, on ne peut obtenir d'estimations fiables pour certains domaines particulièrement petits, les concepteurs d'une enquête devraient envisager sérieusement, de concert avec les promoteurs, la possibilité de regrouper des domaines, de réunir des estimations d'enquêtes différenciées ou encore de ne pas produire d'estimations. Certains domaines ne peuvent être définis à l'avance. Ces domaines non planifiés appellent des méthodes d'estimation spéciales. **Plan de sondage:** Il est rare dans la pratique de trouver un plan "optimal", que ce soit au niveau national ou provincial ou pour un domaine en particulier. Le plus souvent, on introduira des compromis à différentes étapes de l'échantillonnage et de la collecte des données pour répondre aux contraintes théoriques et opérationnelles. Selon les besoins en données, les estimations pour domaines devraient aussi faire partie intégrante de ces compromis. Nous allons étudier deux façons de tenir compte des besoins en données régionales à l'étape de l'élaboration du plan de sondage, notamment la répartition de l'échantillon et le degré de mise en grappes de l'échantillon.

Méthode de répartition: De façon générale, une méthode de répartition optimale pour des estimations nationales permet de répartir les échantillons dans les provinces au prorata de leur population. La fiabilité des estimations pour les petites provinces en souffre. Il est donc préférable d'effectuer une répartition intermédiaire. Il y a diverses façons d'effectuer cette répartition, selon l'importance que l'on accorde aux estimations provinciales ou aux estimations des États (dans le cas de E.-U.). Dans le cas des grandes provinces, une faible réduction de la taille de l'échantillon aura généralement peu d'effet sur la fiabilité des données pour ces provinces (ou des données nationales), mais une augmentation équivalente de la taille d'échantillon dans les petites provinces aura un effet positif appréciable sur la fiabilité des données relatives à ces provinces.

Le même raisonnement s'applique aux domaines planifiés à l'intérieur des provinces car dans la plupart des cas, les répartitions optimales autorisent un certain "jeu" et les concepteurs peuvent exploiter cette caractéristique en redistribuant l'échantillon par un transfert d'unités des grandes régions vers des domaines planifiés de taille moindre.

Mise en grappes: Les enquêtes-ménages de grande envergure comportent habituellement des plans à plusieurs degrés stratifiés avec des unités primaires d'échantillonnage

relativement grandes afin d'assurer un bon rapport coût-efficacité pour les estimations nationales et provinciales. Ce genre de plans implique donc de nombreuses grappes et nuisent par conséquent à la production de statistiques pour des domaines spatiaux non planifiés en ce sens que, sous l'effet du hasard, certains domaines peuvent contenir de nombreuses unités d'échantillon alors que d'autres peuvent n'en contenir aucune. Etant donné l'importance des estimations pour domaines, on devrait s'efforcer de limiter le plus possible la formation de grappes dans l'échantillon. À cet égard, les facteurs suivants jouent un rôle important: choix de la base de sondage, choix des unités d'échantillonnage et de la taille de ces unités, nombre de strates et leur taille et degrés d'échantillonnage. L'objectif est de réduire au maximum les effets de plan étant donnée les contraintes opérationnelles.

Estimation: Quel que soit le degré d'attention que l'on prête aux estimations pour domaine dans les premières étapes de la planification et de la conception d'une enquête, il existera toujours de petits domaines pour lesquels il faudra appliquer des méthodes d'estimation spéciales afin d'obtenir des valeurs suffisamment justes. Depuis quelque temps, on s'intéresse d'une manière particulière aux estimateurs synthétiques; ceux-ci empruntent de l'information à des domaines qui ressemblent au domaine étudié. Or, comme les estimateurs synthétiques sont très sensibles à l'hypothèse de la similarité des domaines, il en faudrait peu pour que le biais théorique soit élevé et que l'utilisation de ces estimateurs soit remise en question. Préoccupés du biais théorique, les concepteurs de plan de sondage ont proposé de combiner des estimateurs directs et des estimateurs synthétiques dans le but de résoudre le problème du biais théorique tout en essayant de conserver les qualités empiriques de Bayes et de méthodes similaires pour pondérer chaque composante des estimateurs combinés. Nous traiterons brièvement ce sujet dans la section 6, qui porte sur l'estimation.

5. CONSIDÉRATIONS RELATIVES AU PLAN D'ÉCHANTILLONNAGE

5.1 Introduction

En règle générale, on pense résoudre le problème des petits domaines (ou des petites régions) par l'estimation. Or, comme nous l'avons mentionné dans la section précédente, il est possible d'envisager la question dès l'étape de la conception du plan de sondage. Pour illustrer nos propos, nous allons nous servir de l'Enquête sur la population active (EPA) du Canada.

Voici les caractéristiques du plan d'échantillonnage actuel de l'EPA: l'Enquête sur la population active du Canada est une enquête mensuelle menée auprès de 59,000 ménages qui sont échantillonnés en plusieurs étapes par diverses méthodes. L'unité d'échantillonnage ultime, c.-à-d. le ménage, demeure dans l'échantillon pendant six mois, puis est remplacée. Les unités des degrés supérieurs (unités primaires d'échantillonnage (UPÉ), grappes) sont

cas, comme le souligne Fellegi (1987), "(...) nous devrions disposer de procédures de révision rigoureuses et vérifiables afin de nous assurer que nous ne combinons ces données que lorsque le bien public qui résulte de ces nouvelles informations statistiques l'emporte manifestement sur l'intrusion dans la vie privée que leur création entraîne."

4. NÉCESSITÉ D'UNE APPROCHE GLOBALE

Bien que les grandes enquêtes visent surtout à produire des estimations nationales et provinciales, il est rare que les estimations issues de ces enquêtes concernent uniquement les populations nationale ou provinciales dans leur entier. Autrement dit, les grandes enquêtes servent immanquablement à produire des estimations pour divers domaines **recoupés** et, parfois, pour des domaines **spatiaux** (ex.: régions intra-provinciales). Dans beaucoup de cas, on ne cherche pas particulièrement à obtenir un niveau de précision voulu pour le domaine, que ce soit à l'étape de la conception du plan de sondage ou à celle de l'estimation, pourvu que le degré de fiabilité soit (jugé) raisonnable. Les problèmes surgissent lorsqu'un domaine recoupé correspond à une sous-population rare ou qu'un domaine spatial correspond à une petite région, auxquels cas il n'existe aucune estimation ou, si elles existent, les estimations sont de qualité douteuse. Dans un certain nombre de cas, ces problèmes peuvent survenir pour la simple raison qu'on n'a pas été assez attentif à ces besoins au début du processus de planification. Si on doit répondre aux besoins en données régionales à l'aide de données d'enquête, il est nécessaire d'élaborer une approche globale qui visera particulièrement à répondre à ces besoins à chacune des étapes d'une enquête: planification, sondage et estimation. En ce qui regarde le plan de sondage et l'estimation, nous classons les domaines en deux catégories:

Domaines planifiés: Du point de vue de l'échantillonnage, les domaines planifiés sont des strates ou des groupes de strates qui ont été formés en vue de la création d'échantillons. Des exemples de ces domaines au Canada sont les régions intra-provinciales, par exemple les régions économiques, les régions de l'Assurance-chômage et les régions créées aux fins de la planification des services de santé. D'autres exemples de ces domaines seraient les grands comtés, les districts ou les régions intra-provinciales de ce type.

Domaines non planifiés: Ce sont des domaines qui n'ont pas été définis à l'étape de la conception du plan de sondage et qui, par conséquent, peuvent recouper des strates. Ils peuvent avoir n'importe quelle taille et créer des problèmes d'estimation particuliers.

Planification: Comme nous l'avons mentionné plus tôt, les données d'enquêtes périodiques comme l'EPA sont beaucoup plus en demande que les données d'enquêtes ponctuelles. En ce qui concerne les enquêtes périodiques qui sont remanées à tous les cinq ou dix ans, on peut élaborer une stratégie adéquate au moment du remaniement puisque dans ces cas précis, les organismes statistiques sont généralement beaucoup plus en mesure de prévoir les

différente et être beaucoup moins élevés que les c.v. de plan correspondants pour la même petite région et, dans beaucoup de cas, moins élevés que les c.v. de plan pour des régions beaucoup plus grandes.

Il est généralement simple d'établir l'expression de l'erreur quadratique moyenne (c.-à-d., variance théorique + carré du biais théorique) pour les estimateurs de modèle. Il n'en va pas de même lorsqu'il s'agit d'estimer cette expression avec un degré de précision raisonnable. S'il est vrai que les données (ex.: taille d'échantillon) pour de petits domaines ne se prêtent pas à la production d'estimations de plan, il est peu probable qu'elles puissent servir à produire une estimation de la variance et du biais corrects. Comme il est relativement plus difficile d'estimer le biais, certains auteurs recherchent chez les estimateurs de modèle la propriété de convergence selon le plan, ce qui laisse supposer peut-être qu'on peut faire abstraction du biais. Or, si la taille de l'échantillon à l'intérieur d'un domaine est suffisamment grande pour qu'un estimateur de modèle soit convergent, un estimateur de plan devrait produire des estimations fiables pour ce domaine. En ce qui concerne les estimateurs de modèle, on a proposé d'utiliser l'estimation de l'erreur quadratique moyenne calculée pour l'ensemble des domaines. Comme, en règle générale, on cherche à obtenir des estimations pour chacun des domaines pris individuellement parce qu'on croit que ces domaines sont différents les uns des autres, on doit s'efforcer d'expliquer pourquoi les estimations de ces divers domaines ont toutes le même degré de fiabilité. Une autre possibilité serait de calculer des estimations indirectes (basées sur un modèle) de la variance et du biais des estimateurs de modèle pour des domaines **pris individuelle-ment**. De fait, l'élaboration de méthodes qui permettraient d'estimer l'erreur quadratique moyenne pour des domaines pris individuellement devrait figurer parmi les priorités de recherche. Une autre préoccupation majeure des expérimentateurs est de savoir comment parer aux défaillances de modèle. Peut-être faudrait-il effectuer des recherches sur la validation de modèle dans des enquêtes complexes. En outre, dans le cas des estimateurs de modèle qui utilisent des données relatives à la variable étudiée pour d'autres périodes que la période pertinente, les estimations de la **variation** d'une période à l'autre seraient de qualité douteuse; voir Schabale (1992). Par surcroît, les estimateurs de modèle qui empruntent de l'information à d'autres domaines de la région présenteraient la même lacune si on comparait des différences entre deux domaines de la région.

Protection de la vie privée: Afin de créer des bases de données complètes desquelles on peut tirer des statistiques régionales, on doit parfois combiner des données de recensement et d'enquête et des données de fichiers administratifs. Cette opération nécessite le couplage d'enregistrements de diverses sources. Or, compte tenu de l'intérêt qu'attache la population à la protection de la vie privée, on ne devrait effectuer des couplages d'enregistrements qu'après en avoir analysé minutieusement toutes les conséquences. La Loi sur la statistique permet à Statistique Canada d'avoir accès aux fichiers administratifs d'autres ministères ou organismes pour des raisons statistiques. Même dans ce

3. QUESTIONS RELATIVES À L'ESTIMATION POUR DOMAINE

La fourniture de données régionales soulève de nombreuses questions d'ordre stratégique ou technique. L'importance de ces questions peut varier d'un organisme à l'autre et d'une application à l'autre au sein du même organisme, selon la qualité des données et les politiques de diffusion. Ces questions intéressent les estimations nationales et provinciales, mais elles supposent une plus grande importance pour les données régionales. Selon les termes mêmes de Brackstone (1987a), "(...) en ce qui concerne l'évaluation de données régionales, il convient de noter que les utilisateurs peuvent observer plus facilement une erreur dans des estimations régionales que dans des agrégats nationaux (...). Il se trouvera des détecteurs prompts à souligner les lacunes des données régionales (...). Il est vrai qu'en ce qui concerne les petites régions, pour lesquelles l'estimation est plus difficile, les valeurs estimées font l'objet d'un examen plus intense." Deux ouvrages en particulier Platak et coll. (1987) et Platak et Singh (1986) présentent plusieurs études de recherche et d'élaboration sur l'estimation pour petites régions. Pour avoir un aperçu des méthodes d'estimation pour petites régions utilisées actuellement dans les programmes statistiques fédéraux aux E.-U., prière de se référer au document du U.S. Statistical Policy Office (1993).

Utilisation de fichiers administratifs: Dans la plupart des cas, l'offre et la demande de données régionales sont influencées en tout premier lieu par les politiques des gouvernements fédéral et provinciaux. Du côté de l'offre, les fichiers administratifs basés sur des programmes gouvernementaux renferment une masse de renseignements qui peuvent servir à produire des données régionales. Des exemples de ces fichiers au Canada: allocations familiales, assurance-chômage, impôt sur le revenu, santé, éducation, sécurité de la vieillesse. On produit déjà de façon régulière des données régionales sur le revenu. Tout **changement** dans la politique du gouvernement et les programmes qui s'y rattachent peut avoir un effet immédiat, positif ou négatif, sur le champ d'application, la disponibilité, l'actualité ou la qualité des données tirées des fichiers administratifs correspondants. En ce qui concerne la demande, les administrations publiques ont besoin de données régionales pour la planification, l'administration et la surveillance de leurs programmes.

Problèmes conceptuels: Dans des séries de données, on confond très souvent les problèmes conceptuels ou défini-

prêtent une attention soutenue à ces estimations régionales. Mais cet intérêt a trait plus souvent à des questions **conceptuelles** qu'à des questions d'estimation; on s'intéresse, par exemple, sur la place des travailleurs découverts, des mises à pied ou des licenciements et des techniques de recherche d'emploi dans le questionnaire d'enquête.

Utilisation de modèles et des mesures de qualité correspondantes: Presque toutes les grandes enquêtes ont leurs estimations pour domaines et aucun problème ne se pose tant que les estimateurs de plan, c.-à-d. des estimateurs approximativement non biaisés selon le plan, sont de qualité acceptable. On considère deux catégories d'estimateurs de plan. D'après Schahale (1992), les estimateurs **directs** sont des estimateurs qui utilisent des valeurs de la variable étudiée qui se rapportent uniquement à des unités du domaine pour la période visée (ex.: l'estimateur par régression dont la pente est estimée au moyen de l'échantillon dans le domaine seulement). Les estimateurs de ce genre peuvent aussi utiliser – et cela est fréquent – de l'information relative à une ou à plusieurs variables auxiliaires pour d'autres domaines ou d'autres périodes, et ces estimateurs sont (approximativement) non biaisés selon le plan. Il existe une autre catégorie d'estimateurs, que nous appelons **estimateurs directs modifiés** et qui peuvent utiliser de l'information relative à la variable auxiliaire et à la variable étudiée pour d'autres domaines ou d'autres périodes mais qui n'ont pas nécessairement la propriété d'être non biaisés selon le plan.

La plupart des producteurs et des utilisateurs de données d'enquête sont familiarisés avec les estimateurs de plan et les inférences correspondantes. Ils interprètent les données sur la base d'échantillons répétés formés à l'aide d'un plan d'échantillonnage probabiliste donné et se servent des c.v. de plan estimés (c.v. = coefficients de variation: racine carrée de la variance de plan divisée par la valeur estimée selon le plan) comme mesure de la qualité des données. Dans les cas où les domaines sont trop petits ou bien le plan de sondage ne prévoit pas la production d'estimations régionales, les estimations de plan peuvent donner des c.v. élevés et les estimations de modèle peuvent alors s'imposer comme la seule solution s'il faut produire des estimations d'enquête pour domaines. La tâche qui attend les statisticiens n'est pas simple: comment estimer et comparer la précision relative des estimations d'une enquête qui produit un grand nombre d'estimations nationales et infra-nationales et de nombreuses estimations pour de grands et de petits domaines en utilisant surtout des estimateurs de plan mais aussi, dans une moindre mesure, des estimateurs de modèle, et comment présenter cela aux utilisateurs. Les c.v. de modèle (racine carrée de la variance théorique de la valeur estimée selon le modèle, divisée par la valeur estimée selon le modèle) peuvent traduire une situation totalement

des établissements) produisent des données fiables de tous genres aux niveaux national, provincial et infra-provincial à l'intention des administrations publiques fédérale et provinciales, des institutions privées, des universités et des médias. Comme la planification, l'administration et la surveillance des programmes sociaux et financiers tendent de plus en plus à se faire au niveau régional, on cherche à obtenir un plus grand nombre de données et des données de meilleure qualité à ce niveau. Nous examinons brièvement ci-dessous trois grandes sources de données socio-économiques et démographiques en mettant l'accent sur les statistiques régionales.

Recensement de la population: Le recensement quinquennal de la population produit des données-répères et constitue, à tous les cinq ans, la source d'information la plus précieuse ayant trait aux petites régions et à divers caractères, domaines ou groupes cibles tels que les minorités ethniques, les personnes en état d'incapacité, les jeunes et les autochtones.

Fichiers administratifs: Les fichiers administratifs sont une source de données statistiques de plus en plus importante. Les organismes statistiques puisent largement dans ces fichiers en démographie pour produire des estimations régionales (Schmidt 1952, Verma et Basavarajappa 1987). Pour certains domaines, comme la statistique de l'état civil, les fichiers administratifs sont la seule source d'information qui permet de produire des statistiques à des niveaux d'agrégation variés. Pour d'autres domaines, l'usage des fichiers administratifs dépend de leur capacité d'offrir des données à jour et des données de qualité comparative aux recensements ou aux enquêtes. Les fichiers administratifs ne servent pas uniquement à la production de totalisations directes; ils sont utilisés dans un certain nombre de programmes comme source d'information supplémentaire dans le but d'améliorer la qualité d'estimations d'enquête. Ils servent aussi à la construction de bases de sondage pour des enquêtes, par exemple le Registre des entreprises et le Registre des adresses de logements résidentiels à Statistique Canada.

Comme le recensement de la population, les fichiers administratifs sont caractérisés par des niveaux d'agrégation géographique très détaillés, ce qui en fait une source d'information précieuse pour l'établissement de statistiques régionales. Leur accessibilité est maintenant plus grande et grâce aux progrès technologiques récents, ils deviennent une source de données ayant un meilleur rapport coût-efficacité. Sur le plan du contenu toutefois, les fichiers administratifs ne sont pas aussi complets et les concepts sont définis pour des programmes plutôt qu'à des fins statistiques. Brackstone (1987a, 1987b) donne les détails d'un programme de Statistique Canada qui a pour objet l'élaboration et l'intégration d'un système de fichiers administratifs destiné à produire des données statistiques. Le recueil rédigé par Coombs et Singh (1987) décrit l'expérience d'autres pays en ce qui concerne l'utilisation de fichiers administratifs.

Programme des enquêtes-ménages: Les enquêtes-ménages sont depuis longtemps une source majeure de données économiques et sociales à Statistique Canada.

Les enquêtes de ce programme peuvent être classées en trois catégories: i) l'Enquête sur la population active, ii) les enquêtes spéciales et les enquêtes supplémentaires et iii) les enquêtes longitudinales et cycliques. Nous décrivons brièvement ci-dessous ces enquêtes en indiquant leur rapport avec la question des données régionales.

Ayant été créée comme une enquête trimestrielle en 1945, l'Enquête sur la population active du Canada (EPA) est devenue une enquête mensuelle en 1952. L'information recueillie au moyen de cette enquête s'est diversifiée considérablement au fil des années et aujourd'hui, elle donne un portrait complet et détaillé du marché du travail au Canada. Outre les estimations nationales et provinciales, l'EPA produit régulièrement des estimations pour les régions infra-provinciales. Il existe aussi une forte demande pour des estimations d'indicateurs courants du marché du travail relatives à de petites régions comme les circonscriptions électorales fédérales, les divisions de recensement et les territoires des centres d'emploi du Canada. Ces estimations servent aux administrations fédérale et provinciales pour contrôler les programmes et répartir les ressources, monétaires ou autres, entre les divers champs de compétence politique et administrative.

Pour des considérations de coût, l'appareil de l'EPA est souvent utilisé pour effectuer des enquêtes ponctuelles et des enquêtes périodiques à l'échelle nationale et provinciale; ce peut être sous la forme d'enquêtes supplémentaires ou d'enquêtes spéciales. Dans le cas des enquêtes supplémentaires, on demande aux participants à l'EPA de répondre à des questions additionnelles tandis que dans le cas des enquêtes spéciales, on demande à un autre échantillon de ménages, tiré de la base de l'EPA, de répondre aux questions. En règle générale, les enquêtes spéciales et les enquêtes supplémentaires sont paratraitées par d'autres ministères et réalisées selon le principe de recouvrement des coûts. Dans le cas de ces enquêtes, la demande de statistiques régionales varie beaucoup selon l'enquête et, en règle générale, elle semble moins pressante que celle provenant de l'EPA.

Statistique Canada effectue annuellement une Enquête sociale générale (ESG) pour répondre, modestement, aux besoins grandissants en données sur des sujets d'actualité sociale. Le programme de l'ESG (Norris et Paton 1991) comporte cinq cycles d'enquête chacun ayant une thématique principale qui reviennent tous les cinq ans. À cause de la taille limitée de l'échantillon (10,000 ménages à l'échelle nationale), l'ESG met l'accent sur les estimations nationales et les données analytiques.

Les enquêtes longitudinales (ou enquêtes par panel) sont nouvelles dans le contexte canadien. Statistique Canada a mis sur pied deux enquêtes longitudinales qui viendront enrichir largement le programme des enquêtes-ménages; ce sont l'Enquête sur la dynamique du travail et du revenu et l'Enquête nationale sur la santé de la population. Il s'agit de deux grandes enquêtes par panel dont on s'attend déjà qu'elles produisent des données au niveau infra-provincial et régional.

Les petites régions: problèmes et solutions

M.P. SINGH, J. GAMBINO et H.J. MANTTEL¹

RÉSUMÉ

Dans cet article, nous discutons de questions techniques relatives à la fourniture de données régionales tirées de recensements, de fichiers administratifs et d'enquêtes. Bien qu'il s'agisse de questions d'ordre général, nous les analysons en relation avec des programmes de Statistique Canada. Nous faisons ressortir la nécessité d'élaborer une approche globale pour les estimations d'enquête et nous soulignons, dans la perspective du remaniement d'une enquête-ménages, les aspects de la conception de plans de sondage qui ont une incidence sur les données régionales. Enfin, nous faisons un survol des méthodes d'estimation en faisant ressortir leurs avantages et leurs inconvénients.

MOTS CLÉS: Stratégie de plan d'échantillonnage; estimateurs de plan; estimateurs de modèle.

1. INTRODUCTION

Pendant des décennies, les fichiers administratifs et les recensements étaient les principales sources de données destinées à l'élaboration des politiques et à la planification pour les grandes et les petites régions. Ils constituent encore aujourd'hui la source de données la plus précieuse pour les petites régions dans la plupart des pays. Dans les années quarante et cinquante toutefois, comme les enquêtes par sondage prenaient de l'importance, les estimations d'enquête sont venues compléter les sources habituelles car elles constituent des données plus à jour et d'un bon rapport coût-efficacité pour un grand nombre de domaines spécialisés. Bien qu'elles soient conçues tout d'abord pour produire des estimations fiables aux niveaux d'agrégation supérieurs, comme les niveaux national et provincial, les enquêtes par sondage visent de plus en plus à répondre à la demande croissante d'estimations plus actuelles pour des domaines de types et de tailles variés. Aucun problème technique ne se pose tant que les domaines sont suffisamment grands (ex.: groupes d'âge-sexe, grandes villes et régions infra-provinciales) pour produire des estimations raisonnablement fiables. Cependant, si on a besoin de données pour de petits domaines, particulièrement des domaines dont les éléments sont répartis dans plusieurs strates de plan, des problèmes d'estimation particuliers se posent et plusieurs méthodes, proposées récemment, existent pour les résoudre.

Cet article a principalement pour but de souligner la nécessité d'envisager le problème des données régionales dans sa globalité. On devrait prendre conscience de la question des petites régions dès le début de la conception des plans de sondage pour les grandes enquêtes. Les plans d'échantillonnage devraient être conçus de manière que l'on puisse obtenir des données régionales fiables à l'aide d'estimateurs de plan ou de modèle. La solution voulant que les organismes statistiques réglient cette difficulté grandissante à l'étape de l'estimation ne devrait être envisagée qu'en dernier recours.

2. BESOINS EN INFORMATION ET SOURCES DE DONNÉES

En tant qu'organisme statistique national, Statistique Canada joue un rôle intégral dans la marche de la société canadienne. Tandis que le caractère confidentiel des données fournies par les répondants est préservé, l'information produite par l'organisme décrit les conditions économiques et sociales du pays et de sa population. Les divers programmes de l'organisme (statistique économique, statistique démographique, statistique sociale et statistique

Dans la section 2, nous examinons les besoins en information et les trois principales sources de données socio-économiques au Canada, c'est-à-dire le recensement, les fichiers administratifs et les enquêtes. Dans la section suivante, nous traitons quelques questions techniques concernant les trois sources de données et nous mettons en lumière le problème des mesures de qualité et de leur interprétation. Ensuite, nous faisons ressortir la nécessité d'élaborer une approche globale qui recouvre les trois grandes étapes de la conception d'une enquête: planification, élaboration du plan de sondage et estimation. Nous examinons particulièrement deux aspects du plan de sondage, à savoir la formation de grappes dans un plan à plusieurs degrés et la répartition de l'échantillon. Dans la section 5, nous présentons quelques caractéristiques de plan de sondage qui ont été introduites récemment dans l'enquête sur la population active du Canada la plus grande enquête mensuelle auprès des ménages de Statistique Canada à l'occasion de son remaniement afin qu'elle produise des données régionales de meilleure qualité. Enfin, la section 6 a pour but d'examiner les nombreuses méthodes d'estimation qui s'appliquent aux petites régions. En même temps, nous proposons de nouveaux estimateurs et nous faisons des commentaires sur les avantages et les faiblesses de divers estimateurs pour domaine. Nous conseillons la prudence dans l'utilisation d'estimateurs de modèle.

¹ M.P. Singh, J. Gambino et H.J. Mantel, Statistique Canada, 16^{ème} étage, Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

Dans ce numéro

Ce numéro de *Techniques d'enquête* commence par une section spéciale sur l'estimation pour petites régions. Les trois articles de cette section traitent le problème de l'estimation pour domaine dans des perspectives différentes. Je voudrais remercier tout particulièrement Jon Rao d'avoir coordonné la rédaction de cette section spéciale. Un ou deux autres articles sur le même sujet, qui n'étaient pas prêts au moment de la publication, pourront apparaître dans une prochaine édition.

Le premier article de la série, de Singh, Gambino et Mantel, examine la question des statistiques régionales du point de vue de la conception des plans de sondage. Les auteurs étudient comment certains aspects du plan de sondage, par exemple la stratification, la formation de grappes et la répartition de l'échantillon, influent sur la production de données régionales pour des domaines planifiés ou non planifiés. Ils font aussi un survol des méthodes d'estimation pour petites régions actuellement en usage. L'article est suivi des commentaires pénétrants de MM. Fuller et Kalton et d'une réponse des auteurs. L'article de Holt et Holmes présente une méthode d'estimation pour petites régions fondée sur un modèle qui ne permet pas l'"emprunt d'information" à d'autres domaines et qui peut être utilisée quand on ne possède pas d'information supplémentaire comme des totaux ou des moyennes de population. Les estimations des paramètres de modèle sont combinées avec des estimations (fondées sur un plan) de moyennes ou de totaux de covariables. En se servant d'une étude de marché comme exemple, les auteurs montrent que la méthode proposée peut amener des gains d'efficacité appréciables dans les estimations pour petits domaines. Le dernier article de la section spéciale, de Singh, Mantel et Thomas, présente une comparaison empirique de divers estimateurs pour petites régions faite à l'aide d'un échantillonnage simulé appliqué à une population de fermes. Les auteurs montrent que, pour les enquêtes à passages répétés, les estimateurs fondés sur un modèle chronologique peuvent – du point de vue tant du biais que de l'erreur quadratique moyenne – être plus efficaces que les estimateurs fondés sur un modèle s'appliquant à une seule période. Kovar et Chen présentent les résultats d'une étude de simulation dans laquelle ils ont analysé les propriétés statistiques d'une méthode "jackknife" d'estimation de la variance pour des ensembles de données imputées. Selon cette méthode, la variance due à l'imputation est intégrée dans l'estimateur de la variance. Les auteurs ont recours à des ensembles de données réelles, à quatre méthodes d'imputation, à l'échantillonnage aléatoire simple et à un mécanisme de non-réponse uniforme. Ils étudient aussi l'efficacité de l'estimation suivant un plan à plusieurs degrés stratifié et un mécanisme de non-réponse non uniforme.

Tracy et Osahan proposent des estimateurs par quotient pour deux méthodes d'échantillonnage qui servent à estimer la moyenne d'une population en grappes chevauchantes lorsque la taille de la population est inconnue. De nombreux travaux de recherche ont été faits sur les grappes non chevauchantes, mais, dans la pratique, il y a beaucoup de cas où l'échantillonnage se fait avec des grappes chevauchantes. La première méthode proposée consiste en un échantillonnage avec remise avec probabilités égales, tandis que la seconde consiste en un échantillonnage avec probabilités inégales. Prasad et Grahm étendent la "méthode des groupes aléatoires" pour un échantillonnage avec probabilité proportionnelle à la taille (PPT) à un échantillonnage effectué en deux occasions. À cette fin, ils se servent de l'information relative à une variable observée à la première occasion pour prélever la portion apparue de l'échantillon à la deuxième occasion. Sitter et Skinner montrent comment, dans les problèmes de stratification multidimensionnelle, on peut obtenir un plan de sondage optimal au moyen de la programmation linéaire. Ils comparent leur méthode avec les méthodes existantes, d'une part en examinant les plans de sondage produits pour des applications particulières et, d'autre part, en évaluant les erreurs quadratiques moyennes. L'estimation de la variance est également examinée. Fuller, Loughin et Baker étudient la production de poids de régression en situation de non-réponse. Ils exposent les conditions dans lesquelles l'estimateur par régression demeure convergent en situation de non-réponse et examinent les facteurs qui peuvent déterminer le choix des variables explicatives. Pour illustrer leurs propos, les auteurs se servent de la Nationwide Food Consumption Survey de 1987-1988, réalisée par le Human Nutrition Information Service du Département de l'Agriculture des E.-U. Enfin, l'article de Stasny, Toomey et Furst décrit une enquête effectuée en 1990 dans le but d'estimer le taux de sans-abri dans les régions rurales de l'Ohio. Les auteurs évaluent la taille probable du biais de l'estimateur en simulant un échantillonnage dans diverses populations synthétiques. Ils constatent que le biais sera plutôt négligeable par rapport à l'écart type.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 20, numéro 1, juin 1994

TABLE DES MATIÈRES

1	Dans ce numéro
	Estimation pour petites régions
3	M.P. SINGH, J. GAMBINO et H.J. MANTEL
16	Les petites régions: problèmes et solutions
19	Commentaires: W.A. FULLER
22	G. KALTON
	Réponse des auteurs
25	D. HOLT et D.J. HOLMES
	Estimation pour petits domaines dans des plans de sondage avec probabilités
	inégaies
35	A.C. SINGH, H.J. MANTEL et B.W. THOMAS
	MPLSE à données chronologiques pour petites régions évalués à l'aide de
	données d'enquête
47	J.G. KOVAR et E.J. CHEN
	Méthode du jackknife pour l'estimation de la variance en présence de données
	imputées
57	D.S. TRACY et S.S. OSAHAN
	Estimation pour grappes chevauchantes lorsque la taille de la population est
	inconnue
63	N.G.N. PRASAD et J.E. GRAHAM
	Echantillonnage avec PPT en deux occasions
69	R.R. SITTER et C.J. SKINNER
	Stratification multidimensionnelle par programmation linéaire
79	W.A. FULLER, M.M. LOUGHIN et H.D. BAKER
	Production de poids de régression en situation de non-réponse et application à la
	Nationwide Food Consumption Survey de 1987-1988
91	E.A. STASNY, B.G. TOOMEY et R.J. FIRST
	Estimation du taux de sans-abri dans les régions rurales: une étude des régions
	non urbaines de l'Ohio

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinappa

G.J.C. Hole

F. Mayda (Directeur de la Production)

R. Platek (Ancien président)

M.P. Singh

D. Roy

C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, Statistique Canada

Rédacteurs associés

D.R. Bellhouse, University of Western Ontario

D. Binder, Statistique Canada

M.J. Colledge, Statistique Canada

J.-C. Deville, INSEE

J.D. Drew, Statistique Canada

W.A. Fuller, Iowa State University

M. Gonzalez, U.S. Office of Management and Budget

R.M. Groves, U.S. Bureau of the Census

D. Holt, University of Southampton

G. Kalton, University of Michigan

A. Mason, East-West Center

Rédacteurs adjoints

N. Laniel, M. Latouche, L. Mach et H. Mantel, Statistique Canada

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

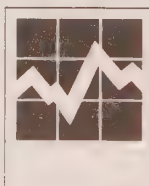
Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001 au catalogue) est de 45 \$ par année au Canada, 50 \$ (E.-U.) aux États-Unis, et de 55 \$ (E.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

TECHNIQUES D'ENQUÊTE



UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

JUIN 1994 • VOLUME 20 • NUMÉRO 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministère de l'Industrie, des Sciences
et de la Technologie, 1994

Tous droits réservés. Il est interdit de reproduire ou de transmettre
le contenu de la présente publication, sous quelque forme ou
par quelque moyen que ce soit, enregistré ou non, sur support
magnétique, reproduction électronique, mécanique, photographique,
ou autre, ou de l'emmagasiner dans un système de recouvrement,
sans l'autorisation écrite préalable des Services de concession
des droits de licence, Division de la commercialisation,
Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 1994

Prix : Canada : 45 \$

États-Unis : 50 \$ US

Autres pays : 55 \$ US

N° 12-001 au catalogue

ISSN 0714-0045

Ottawa





NUMÉRO 1

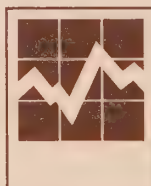
VOLUME 20

Juin 1994

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

Catalogue 12-001

TECHNIQUES D'ENQUÊTE



12
-001



Gover
Publication

SURVEY METHODOLOGY

Catalogue 12-001

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1994

VOLUME 20

NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1994 • VOLUME 20 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1994

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

December 1994

Price: Canada: \$45.00
United States: US\$50.00
Other Countries: US\$55.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members B.N. Chinnappa C. Patrick
G.J.C. Hole D. Roy
F. Mayda (Production Manager) M.P. Singh
R. Platek (Past Chairman)

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
M.J. Colledge, <i>Australian Bureau of Statistics</i>	L.-P. Rivest, <i>Université Laval</i>
J.-C. Deville, <i>INSEE</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J.D. Drew, <i>Statistics Canada</i>	C.-E. Särndal, <i>Université de Montréal</i>
J.-J. Droesbeke, <i>Université Libre de Bruxelles</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
W.A. Fuller, <i>Iowa State University</i>	F.J. Scheuren, <i>George Washington University</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	J. Sedransk, <i>State University of New York</i>
R.M. Groves, <i>University of Maryland</i>	J. Waite, <i>U.S. Bureau of the Census</i>
D. Holt, <i>University of Southampton</i>	J. Waksberg, <i>Westat, Inc.</i>
G. Kalton, <i>Westat, Inc.</i>	K.M. Wolter, <i>National Opinion Research Center</i>
A. Mason, <i>East-West Center</i>	A. Zaslavsky, <i>Harvard University</i>

Assistant Editors N. Laniel, M. Latouche, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue No. 12-001) is \$45 per year in Canada, US \$50 in the United States, and US \$55 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 20, Number 2, December 1994

CONTENTS

In This Issue	95
---------------------	----

Establishment Survey Methods

J. ARMSTRONG and H. ST-JEAN Generalized Regression Estimation for a Two-Phase Sample of Tax Records	97
--	----

F.J. GALLEGO, J. DELINCÉ and E. CARFAGNA Two-Stage Area Frame Sampling on Square Segments for Farm Surveys	107
---	-----

K.H. POLLOCK, S.C. TURNER and C.A. BROWN Use of Capture-Recapture Techniques to Estimate Population Size and Population Totals when a Complete Frame is Unavailable	117
---	-----

A.R. GOWER Questionnaire Design for Business Surveys	125
---	-----

E. RANCOURT, H. LEE and C.-E. SÄRNDAL Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse	137
--	-----

Y. DING and S.E. FIENBERG Dual System Estimation of Census Undercount in the Presence of Matching Error	149
---	-----

P.S. KOTT A Hypothesis Test of Linear Regression Coefficients with Survey Data	159
---	-----

L.H. COX Matrix Masking Methods for Disclosure Limitation in Microdata	165
---	-----

P.D. FALORSI, S. FALORSI and A. RUSSO Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey	171
---	-----

T. NIYONSENGA Nonparametric Estimation of Response Probabilities in Sampling Theory	177
--	-----

O. SCHABENBERGER and T.G. GREGOIRE Competitors to Genuine π ps Sample Designs: A Comparison	185
--	-----

Acknowledgements	193
------------------------	-----

In This Issue

This issue of *Survey Methodology* opens with a special section on **Establishment Survey Methods**. The four papers in this special section deal with important issues in the context of establishment surveys such as questionnaire design, sample design and estimation. These papers were initially presented at the International Conference on Establishment Surveys, Buffalo, New York, June 1993.

The paper by Armstrong and St-Jean presents an application of the general framework of regression estimation in two-phase sampling. Using data from a two-phase sample of tax records, three particular cases of the generalized regression estimator – two regression estimators and a poststratified estimator – are empirically compared to the Horvitz-Thompson estimator. The empirical study shows that the poststratified estimator is more efficient than the Horvitz-Thompson estimator and as efficient as the two regression estimators.

Gallego, Delincé and Carfagna describe the Monitoring Agriculture with Remote Sensing (MARS) project of the European Community. As the project is not capable of producing good estimates of crop areas and yields, they describe a method of sampling farms by points to obtain reliable estimates. Results of applying this approach in two regions, Emilia Romagna in Italy and the Czech Republic, are described.

Pollock, Turner and Brown discuss the use of capture-recapture sampling to estimate the population size and population totals when only incomplete list frames exist. A discussion of the properties of the resulting model based estimators and an example where the establishments are fishing boats are presented.

In the last paper of this special section, Gower gives an overview of important considerations that should be taken into account when developing and designing questionnaires for business surveys. Examples of applications of focus groups and cognitive research to test questionnaires for business surveys are presented.

Rancourt, Lee and Särndal present simple correction factors to reduce the bias of the standard estimator of the population mean in the case of ratio imputation for confounded nonresponse. The effectiveness of these factors is studied by Monte Carlo simulations. The factors are found to be effective especially when the model underlying ratio imputation holds.

The use of the capture-recapture approach for coverage evaluation of the U.S. census is discussed by Ding and Fienberg. They give methods for estimating population total and census undercount when the assumption of a perfect match between individuals in the census and in the sample is relaxed. They propose models to describe two types of matching errors, mismatches and erroneous non-matches. The methods are illustrated using data from 1986 Los Angeles test census and 1990 Decennial Census.

Kott discusses testing a hypothesis about linear regression coefficients using data from a sample survey. He suggests an adjustment of the design-based linearization variance estimator to reduce its model bias and a formula to estimate its effective degrees of freedom. Two examples of the method are presented.

Cox develops a framework, called matrix masking, for microdata disclosure limitation methods that should improve understanding of these methods and of their effect on data use. Within this framework, based on ordinary matrix arithmetic, statistical agencies can develop, evaluate and use reliable software for disclosure limitation of microdata. The author presents explicit matrix mask formulations for the principal microdata masking methods in current use.

Falorsi, Falorsi and Russo conduct an empirical comparison of some small area estimation methods in the context of the Italian Labour Force Survey using data from the 1981 Italian Census. The estimators included in their study are a poststratified direct estimator, a synthetic estimator, an optimal linear combination of the two, and a sample size dependent estimator. They conclude that, for their application, the sample size dependent estimator offers the best balance of variance and bias.

The paper by Niyonsenga presents a comparison of two nonparametric methods of estimation of response probabilities in sampling theory via a Monte Carlo simulation. It is shown that, in the context of simple random sampling without replacement, the nonparametric variant based on the ranks of the values of the auxiliary variable performs better, with respect to both bias and mean square error, than the method based on the values of the auxiliary variable, for both the expansion and regression estimators.

Schabenberger and Gregoire compare alternative exact and approximate π ps strategies in the context of sampling in forestry. Two sequential sampling schemes due to Sunter combined with the Horvitz-Thompson estimator are compared to the random group strategy of Rao, Hartley and Cochran (RHC) as well as a ratio of means estimator used with simple random sampling. If the size variable is highly correlated with the variable of interest then π ps strategies are considerably more efficient. When the correlation is very high the exact π ps strategy is most efficient; however, the RHC strategy has the advantage of simplicity. If the correlation is low then the π ps strategies can be very inefficient.

The Editor

Generalized Regression Estimation for a Two-Phase Sample of Tax Records

JOHN ARMSTRONG and HÉLÈNE ST-JEAN¹

ABSTRACT

A generalized regression estimator for domains and an approximate estimator of its variance are derived under two-phase sampling for stratification with Poisson selection at each phase. The derivations represent an application of the general framework for regression estimation for two-phase sampling developed by Särndal and Swensson (1987) and Särndal, Swensson and Wretman (1992). The empirical efficiency of the generalized regression estimator is examined using data from Statistics Canada's annual two-phase sample of tax records. Three particular cases of the generalized regression estimator – two regression estimators and a poststratified estimator – are compared to the Horvitz-Thompson estimator.

KEY WORDS: Model assisted estimation; Domain estimation; Poisson sampling.

1. INTRODUCTION

In this paper the problem of domain estimation under two-phase sampling for stratification is examined in a case in which Poisson sampling is used at both phases of selection. Consider a population of N units and suppose that it is necessary to estimate the total of a characteristic of interest, y , for L disjoint domains. Domain membership can be well, but not exactly, predicted using an auxiliary variable, θ , that is not observed before sampling. The cost of obtaining information on θ is lower than the cost of obtaining information on y and lower than the cost of obtaining exact domain membership data. At the first phase of sampling, a Poisson sample is drawn from the population and the value of θ is observed for each sampled unit. The units in the first-phase sample are stratified using θ -values. This stratification is an approximation to stratification by domain. At the second phase of sampling, a Poisson sample is drawn from each stratum. The value of y , as well as exact domain membership data, is observed for each unit in the second-phase sample.

The Horvitz-Thompson estimator of the total of y for domain d is $\hat{Y}_{H-T}(d) = \sum_{i \in s_2} y_i(d) / (p_{1i}p_{2i})$, where $y_i(d)$ takes the value of y_i if unit i falls in domain d and otherwise takes the value zero, s_2 denotes the second-phase sample and p_{1i} and p_{2i} are first- and second-phase selection probabilities, respectively, for unit i . Since the sample sizes obtained using Poisson sampling are random variables, this estimator may be inefficient. (See Sunter 1986 or Särndal, Swensson and Wretman 1992, p. 63.) Generalized regression estimation is an alternative to the Horvitz-Thompson estimator that can be employed when auxiliary information is available. A generalized regression

estimator for two-phase Poisson sampling and an approximate estimator of its variance are derived in this paper.

Section 2 contains the derivation of the generalized regression estimator and approximate variance estimator. Section 3 includes a description of the application that motivated the estimation problem – Statistics Canada's annual two-phase sample of tax records. The results of an empirical study comparing the Horvitz-Thompson estimator with three particular cases of the generalized regression estimator – the poststratified estimator currently used in production and two regression estimators – are described in Section 4.

2. GENERALIZED REGRESSION ESTIMATION

Generalized regression estimation is not a new technique. A generalized regression estimator for a one-phase sample design is described by Deming and Stephan (1940). Recent applications of generalized regression estimation at Statistics Canada include the work of Lemaître and Dufour (1987) and Bankier, Rathwell and Majkowski (1992). Hidioglou, Särndal and Binder (1993) provide an extensive discussion of the use of generalized regression estimators for business surveys.

Derivation of generalized regression estimators can be approached from the perspective of model assisted survey sampling (Särndal, Swensson and Wretman 1992) or from the perspective of calibration (Deville and Särndal 1992). Let $U = \{u\}$ and $V = \{v\}$ denote sets of first-phase poststrata and second-phase poststrata, respectively. During generalized regression weighting of the first-phase sample, the design weights $1/p_{1i}$ are adjusted to yield weights $w_{1i} = g_{1i}/p_{1i}$ that respect the calibration equations

¹ John Armstrong, Social and Economic Studies Division, 24 – R.H. Coats Bldg., and Hélène St-Jean, Business Survey Methods Division, 11 – R.H. Coats Bldg., Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

$$\sum_{i \in s1 \cap v} w_{1i} x_i = X_u,$$

for each first-phase poststratum u , where x_i is an $L_1 \times 1$ vector of auxiliary variables known for all units in the population and X_u is the vector of auxiliary variable totals for poststratum u . The adjusted weights minimize the distance measure $\sum_{i \in s1} (g_{1i} - 1)^2 / p_{1i}$. The same weights can be obtained from a model assisted perspective using

$$E_{\xi}(y_i) = x_i' \beta_u, i \in u$$

$$V_{\xi}(y_i) = \sigma^2,$$

where y_i is the value of the variable of interest for unit i , and $E_{\xi}(\cdot)$ and $V_{\xi}(\cdot)$ denote expectation and variance, respectively, with respect to the model.

For the generalized regression estimators of interest, weighting of the second-phase sample involves a calibration procedure that is conditional on the results of first-phase weighting. The initial weights, w_{1i}/p_{2i} , are adjusted to give final weights, $w_i = g_{2i} w_{1i}/p_{2i}$, that satisfy the calibration equations

$$\sum_{i \in s2 \cap v} w_i z_i = \bar{Z}_v,$$

for each second-phase poststratum v , where z_i is an $L_2 \times 1$ vector of auxiliary variables known for all units in the first-phase sample and $\bar{Z}_v = \sum_{i \in s1 \cap v} w_{1i} z_i$ is an estimate of the vector of auxiliary variable totals for post-stratum v , computed using the adjusted first-phase weights w_{1i} . Note that these calibration equations differ in an important way from the examples given by Särndal and Swensson (1987, pp. 284-288) and Särndal, Swensson and Wretman (1992, pp. 359-366) because they involve adjusted first-phase weights rather than first-phase design weights. The final weights minimize the distance measure $\sum_{i \in s2} w_{1i} (g_{2i} - 1)^2 / p_{2i}$. The model needed to obtain the same weights from a model assisted perspective is

$$E_{\xi}(w_{1i} y_i) = w_{1i} z_i' \beta_v, i \in v$$

$$V_{\xi}(w_{1i} y_i) = w_{1i} \sigma^2.$$

Use of adjusted first-phase weights rather than first-phase design weights in the second-phase calibration equations has two important advantages. First, the generalized regression estimator for domain d can be written as

$$\hat{Y}_{\text{GREG}}(d) = \sum_{i \in s2} y_i(d) g_{1i} g_{2i} / p_{1i} p_{2i},$$

using first-phase and second-phase g -weights. Second, suppose that some auxiliary variables are used for calibration at both phases of weighting. Estimates of population totals for such variables that are equal to actual totals can be constructed using final weights.

Let $\bar{X}_u = \sum_{i \in s1 \cap u} x_i / p_{1i}$ denote the $L_1 \times 1$ vector of Horvitz-Thompson estimates of auxiliary variable totals for first-phase poststratum u . The first-phase g -weight is

$$g_{1i} = 1 + \lambda'_u x_i,$$

where $\lambda'_u = (X_u - \bar{X}_u)' M_u^{-1}$ and $M_u^{-1} = (\sum_{i \in s1 \cap u} x_i x_i' / p_{1i})^{-1}$. For second-phase poststratum v , denote the estimate of \bar{Z}_v based on initial second-phase weights by $\bar{Z}_v = \sum_{i \in s2 \cap v} w_{1i} z_i / p_{2i}$. The second-phase g -weight is

$$g_{2i} = 1 + \lambda'_v z_i,$$

where $\lambda'_v = (\bar{Z}_v - \bar{Z}_v)' M_v^{-1}$ and $M_v^{-1} = (\sum_{i \in s2 \cap v} w_{1i} z_i z_i' / p_{2i})^{-1}$.

The approximate variance of $\hat{Y}_{\text{GREG}}(d)$ is given by

$$V(\hat{Y}_{\text{GREG}}(d)) \approx \sum_i \frac{1 - p_{1i}}{p_{1i}} Q_{1i}^2 + E_1 \left[\sum_{i \in s2} \frac{1 - p_{2i}}{p_{2i}} (w_{1i} Q_{2i})^2 \right],$$

where $E_1(\cdot)$ denotes expectation with respect to the first phase of sampling, $Q_{1i} = y_i(d) - x_i' B_u$ for each unit in first-phase poststratum u , and B_u , the vector of estimated coefficients from the regression of $y(d)$ on x that would be obtained if $y(d)$ was available for all units in first-phase poststratum u , is given by

$$B_u = \left(\sum_{i \in u} x_i x_i' \right)^{-1} \left(\sum_{i \in u} x_i y_i(d) \right).$$

Similarly, $Q_{2i} = y_i(d) - z_i' B_v$ for each unit in second-phase poststratum v and B_v , the vector of estimated coefficients from the regression of $y(d)$ on z that would be obtained, conditional on the first-phase calibration, if $y(d)$ was available for all units in the component of the first-phase sample falling in second-phase poststratum v , is given by

$$B_v = \left(\sum_{i \in s1 \cap v} w_{1i} z_i z_i' \right)^{-1} \left(\sum_{i \in s1 \cap v} w_{1i} z_i y_i(d) \right).$$

An estimator of the approximate variance of $\hat{Y}_{\text{GREG}}(d)$ is

$$\hat{V}(\hat{Y}_{\text{GREG}}(d)) = \sum_i \frac{1 - p_{1i}}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2 + \sum_i \frac{1 - p_{2i}}{(p_{1i} p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Since $y(d)$ is available only for units in s_2 , estimates of B_u and B_v are

$$\hat{B}_u = \left(\sum_{i \in s_2 \cap u} w_i x_i x_i' \right)^{-1} \left(\sum_{i \in s_2 \cap u} w_i x_i y_i(d) \right),$$

$$\hat{B}_v = \left(\sum_{i \in s_2 \cap v} w_i z_i z_i' \right)^{-1} \left(\sum_{i \in s_2 \cap v} w_i z_i y_i(d) \right).$$

The sample residuals needed to compute the variance estimator are $q_{1i} = y_i(d) - x_i' \hat{B}_u$ and $q_{2i} = y_i(d) - z_i' \hat{B}_v$. More details of the derivation of the approximate variance of $\hat{Y}_{\text{GREG}}(d)$ and the estimator of the approximate variance are given in Appendix A.

If y is strongly correlated with x and z , the variance of the generalized regression estimator of the population total of y will be relatively small. However, it is important to note that strong correlations between y and x and z will not necessarily lead to a relatively small variance for the estimate of the total of y for a particular domain, since $y(d)$ may be poorly correlated with x and z within poststrata that include at least one sampled unit falling in domain d .

The correlation between $y(d)$ and x and z within a poststratum that includes at least one sampled unit falling in domain d may be low if some sampled units in the poststratum do not fall in domain d . This situation may arise often if domain totals of auxiliary variables and/or exact domain membership information for units in the first-phase sample are unavailable. In the context of two-phase sampling for stratification, there is no domain membership information available before selection of the first-phase sample. If each first-phase poststratum is formed by combining one or more first-phase sampling strata, for example, most first-phase poststrata will include more than one domain. The variable Θ used to predict domain membership during stratification of the first-phase sample is not an exact predictor. If second-phase poststrata are formed by combining second-phase sampling strata, each domain may be divided between a number of second-phase poststrata.

Depending on the type of auxiliary information used, the g -weights associated with the generalized regression estimator and, consequently, generalized regression estimates, may be negative.

3. APPLICATION: TWO-PHASE SAMPLING OF TAX RECORDS

The two-phase tax sample is part of a general strategy at Statistics Canada for production of annual estimates of Canadian economic activity. Annual economic data for

large businesses are collected through mail-out sample surveys. Data for small businesses are obtained from the tax sample. Estimates of financial variables for the business population are obtained by combining tax and survey estimates. Tax data rather than survey data are used to obtain small business estimates in order to reduce costs and response burden.

The two-phase sample design was introduced in response to a requirement for estimates for domains defined using the four-digit Standard Industrial Classification (SIC) code (Statistics Canada 1980). The first two digits of SIC (SIC2) provides a classification of businesses activity into 76 groups. Within each group, four-digit SIC (SIC4) codes provide classification into finer categories. For example, the SIC2 code of a business might classify it in the transportation industry while the SIC4 code describes the activity of the business as bulk liquids trucking.

There are two types of taxfilers – T1s and T2s. A T1 taxfiler is an individual, who may own all or part of one or more unincorporated businesses, while a T2 taxfiler is an incorporated business. Administrative files that contain limited information for all taxfilers that are associated with businesses are provided to Statistics Canada by Revenue Canada, the Canadian government department responsible for tax collection. These files are used to construct a sampling frame. Information concerning numbers of businesses owned by T1 taxfilers and ownership shares is not available on the sampling frame. Frame data does include geographical information, as well as gross business income and net profit for both T1 and T2 taxfilers. A few other major financial variables, including salary and inventory data, are generally available for T2 taxfilers. Estimates are required for about 35 financial variables that can be obtained from tax returns and associated financial statements but are not available on administrative files supplied by Revenue Canada.

Taxfilers that are associated with businesses are classified by Revenue Canada using the SIC system. In most cases, descriptions of business activity reported on tax returns are sufficient to accurately determine SIC2 codes. Revenue Canada assigns additional digits of SIC to most taxfilers. However, not all taxfilers are classified to the four-digit level and the third and fourth digits of SIC4 codes assigned by Revenue Canada are relatively inaccurate. A two-phase approach to sampling of tax records was adopted to facilitate accurate estimation of economic production at the SIC4 level.

Section 3.1 includes a brief description of the two-phase sampling design. More information about the two-phase design is provided in Armstrong, Block and Srinath (1993). Sections 3.2 and 3.3 contain information concerning estimation for the two-phase design. The Horvitz-Thompson estimator is described in Section 3.2 and a poststratified estimator is discussed in Section 3.3.

3.1 Sampling Design

The administrative information used to construct the sampling frame for a particular tax year is accumulated by Revenue Canada over a period of two calendar years as tax returns are received and processed. The use of Poisson sampling offers substantial operational advantages because sampling operations can begin before a complete sampling frame is available.

The target (in-scope) population for tax sampling is the population of businesses with gross income over \$25,000, excluding large businesses covered by mail-out sample surveys. The first-phase sample is a longitudinal sample of taxfilers. Strata are defined by SIC2, province and size (gross business income). All taxfilers that are included in the first-phase sample for tax year T and are still in-scope for tax sampling in tax year $T + 1$ remain in the first-phase sample for tax year $T + 1$. Taxfilers may be added to the first-phase sample each year to improve the precision of certain estimates and to replace taxfilers sampled in previous years that are no longer in-scope.

To implement Poisson sampling for first-phase sample selection, each taxfiler is assigned a pseudo-random number (hash number) in the interval $(0,1)$ generated by a hashing function that uses the unique taxfiler identifier as input. The hash number for each taxfiler is compared to the sampling interval for the corresponding stratum. If the hash number for a particular taxfiler falls in the corresponding sampling interval and the taxfiler is not already in the first-phase sample, then the taxfiler is added to the first-phase sample. Since taxfiler identifiers do not change over time, Poisson sampling facilitates selection of a longitudinal first-phase sample.

First-phase selection probabilities for taxfilers that are already included in the first-phase sample are updated each year. Longitudinal updating is necessary because: (i) a taxfiler may fall in different first-phase sampling strata in consecutive tax years; and (ii) first-phase sampling fractions for a given stratum may vary from one year to the next.

Copies of tax returns and associated financial statements for taxfilers in the first-phase sample are sent to Statistics Canada from Revenue Canada. In order to select the second-phase sample, statistical entities are created using information about businesses corresponding to taxfilers in the first-phase sample. Let $J = \{j\}$ denote the population of businesses that is the target population for tax sampling. A statistical entity, denoted by (i,j) , is created for every taxfiler-business combination in the first-phase sample. For each T1 taxfiler in the first-phase sample, data for all businesses wholly or partially owned by the taxfiler (including ownership shares) that are needed to create statistical entities are available from tax returns and associated financial statements. Since there is a one-to-one correspondence between businesses and T2 taxfilers, a single statistical entity is created for each T2 taxfiler in the first-phase sample.

For each tax year, statistical entities that have not appeared in previous tax samples are assigned SIC4 codes by Statistics Canada. These codes are determined using information supplementary to business activity descriptions reported on tax returns and are more accurate in digits three and four than codes assigned by Revenue Canada. For statistical entities that have appeared in previous tax samples, the SIC4 assigned earlier is carried forward.

Conceptually, the second-phase sample is a sample of businesses. Operationally, it is a sample of taxfilers selected using statistical entities. Statistical entities are stratified using SIC4 codes assigned by Statistics Canada, as well as province and size. The total revenue of business j is used as the size variable for statistical entity (i,j) . If one statistical entity corresponding to a T1 taxfiler is selected for the second-phase sample, then all statistical entities corresponding to the taxfiler are selected. Consequently, the second-phase selection probability for statistical entity (i,j) depends only on i .

Second-phase sample selection is done by the Poisson sampling method using hash numbers generated from taxfiler identifiers. The hashing function used for second-phase sample selection is independent of the first-phase hashing function.

Data for about 35 financial variables are transcribed from tax returns and associated financial statements for taxfilers selected in the second-phase sample. SIC4 codes assigned by Statistics Canada are updated, if necessary, to ensure that all SIC4 codes used during tabulation of estimates correspond to the current tax year.

3.2 Horvitz-Thompson Estimator

The second-phase sample is a sample of businesses selected using statistical entities. Since some businesses are partnerships, more than one statistical entity may correspond to the same business. To construct estimates for the population of businesses, an adjustment for the effects of partnerships is required. If business j is a partnership, it will be included in the second-phase sample if any of the corresponding taxfilers are selected. The usual Horvitz-Thompson estimator must be adjusted for partnerships to avoid over-estimation. Let δ_{ij} denote the proportion of business j owned by taxfiler i and suppose that statistical entity (i,j) is selected for the second-phase sample. The data for business j is adjusted by multiplying it by δ_{ij} so that only the component of income and expense items corresponding to taxfiler i is included in estimates. Rao (1968a) describes a similar adjustment in a slightly different context.

Let y_j denote the value of the variable y for business j . The Horvitz-Thompson estimate of the total of y over domain d , incorporating adjustment for partnerships, is given by

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} \sum_{j \in J_i} \delta_{ij} y_j(d) / (p_{1i} p_{2i}),$$

where J_i is a set containing the indices of the businesses wholly or partially owned by taxfiler i . Since selection probabilities depend only on the taxfiler index i , $\hat{Y}_{H-T}(d)$ can be written as

$$\hat{Y}_{H-T}(d) = \sum_{i \in s2} y_i(d) / (p_{1i} p_{2i}),$$

where

$$y_i(d) = \sum_{j \in J_i} \delta_{ij} y_j(d).$$

$\hat{Y}_{H-T}(d)$ is an unbiased estimator of the population total of y for businesses in domain d . Refer to Rao (1968a).

The second-phase sample is obtained by Poisson sub-sampling of the first-phase Poisson sample. Consequently, the second-phase sample is also a Poisson sample and the variance of $\hat{Y}_{H-T}(d)$ is

$$V(\hat{Y}_{H-T}(d)) = \sum_i [(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})] y_i(d)^2.$$

An unbiased estimator of this variance is

$$\hat{V}(\hat{Y}_{H-T}(d)) = \sum_{i \in s2} [(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})^2] y_i(d)^2.$$

3.3 Poststratified Horvitz-Thompson Estimator

Adjustment of the Horvitz-Thompson estimator to account for differences between actual and expected sample sizes under Poisson sampling was suggested by Brewer, Early and Joyce (1972). The methodology currently used to produce estimates based on the two-phase tax sample incorporates such adjustments.

Ratio adjustments are applied within poststrata during weighting of both the first- and second-phase samples. Choudhry, Lavallée and Hidirolou (1989) provide a general discussion of weighting for a two-phase Poisson sample using poststratified ratio adjustments. Suppose that first-phase poststratum u contains N_u taxfilers. An estimate of the number of taxfilers in the population that fall in first-phase poststratum u , based on the first-phase sample, is

$$\tilde{N}_u = \sum_{i \in s1 \cap u} (1/p_{1i}).$$

The poststratified first-phase weight for taxfiler i , $i \in u$ is

$$w_{1i} = (1/p_{1i}) (N_u / \tilde{N}_u).$$

An estimate of the number of taxfilers in second-phase poststratum v , based on the first-phase sample, is

$$\tilde{N}_v = \sum_{i \in s1 \cap v} w_{1i}.$$

An alternative estimate, using only units in the second-phase sample, is

$$\dot{N}_v = \sum_{i \in s2 \cap v} w_{1i} / p_{2i}.$$

The poststratified second-phase weight for statistical entity (i, j) in poststratum v is

$$w_{2i} = (1/p_{2i}) (\tilde{N}_v / \dot{N}_v)$$

and the final weight is

$$w_i = w_{1i} w_{2i}.$$

The poststratified estimate of the total of y over domain d is

$$\hat{Y}(d) = \sum_{i \in s2} w_i y_i(d).$$

Choudhry, Lavallée and Hidirolou (1989) note that the variance of $\hat{Y}(d)$ is approximately given by

$$\begin{aligned} V(\hat{Y}(d)) \approx & \sum_u \sum_{i \in u} \frac{(1 - p_{1i})}{p_{1i}} \left(y_i(d) - \frac{Y_u(d)}{N_u} \right)^2 \\ & + \sum_v \sum_{i \in v} \frac{(1 - p_{2i})}{p_{1i} p_{2i}} \left(y_i(d) - \frac{Y_v(d)}{N_v} \right)^2, \end{aligned}$$

where $Y_u(d)$ and $Y_v(d)$ are population totals for the variable y over the portions of the domain d belonging to poststrata u and v respectively.

This variance is estimated by

$$\begin{aligned} \hat{V}(\hat{Y}(d)) = & \sum_u \sum_v \left(\frac{N_u}{\tilde{N}_u} \right)^2 \left(\frac{\tilde{N}_v}{\dot{N}_v} \right)^2 \\ & \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} \left(y_i(d) - \frac{\hat{Y}_u(d)}{\hat{N}_u} \right)^2 \\ & + \sum_u \sum_v \left(\frac{N_u}{\tilde{N}_u} \right)^2 \left(\frac{\tilde{N}_v}{\dot{N}_v} \right)^2 \\ & \sum_{i \in s2 \cap u \cap v} \frac{(1 - p_{2i})}{(p_{1i} p_{2i})^2} \left(y_i(d) - \frac{\hat{Y}_v(d)}{\hat{N}_v} \right)^2, \end{aligned}$$

where the estimates \hat{N}_u and \hat{N}_v are calculated using final weights.

The inclusion of the factor $(N_u/\tilde{N}_u)^2(\tilde{N}_v/\tilde{N}_v)^2$ can be motivated by an improvement in the conditional properties of the estimator (Royall and Eberhardt 1975). A variance estimator for the ratio estimator for a one-phase sample design including an analogous adjustment factor has also been studied by Wu (1982). Empirical work reported by Wu and Deng (1983) indicates that the coverage properties of confidence intervals based on the normal approximation are improved using the adjustment factor.

$\hat{Y}(d)$ is a particular case of $\hat{Y}_{\text{GREG}}(d)$ that can be obtained if a single auxiliary variable with value one for all taxfilers is employed during both first- and second-phase weighting. In this case, we have $g_{1i} = N_u/\tilde{N}_u$ for all taxfilers in first-phase poststratum u and $g_{2i} = \tilde{N}_v/\tilde{N}_v$ for all taxfilers in second-phase poststratum v . Note that negative g -weights are precluded by this choice of auxiliary variables. The variance estimator $\hat{V}(\hat{Y}(d))$ differs in a minor way from the estimator $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ for this particular case of $\hat{Y}_{\text{GREG}}(d)$. The second-phase g -weight appears in the leading term of $\hat{V}(\hat{Y}(d))$ but does not appear in $\hat{V}(\hat{Y}_{\text{GREG}}(d))$.

4. EMPIRICAL STUDY

In order to compare the performance of $\hat{Y}_{H-T}(d)$, $\hat{Y}(d)$ and $\hat{Y}_{\text{GREG}}(d)$, an empirical study was conducted using data from the province of Quebec for tax year 1989. Since the estimator $\hat{Y}(d)$ is a special case of $\hat{Y}_{\text{GREG}}(d)$, it will be called $\hat{Y}_{\text{GREG-TPH}}(d)$ in subsequent discussion. (TPH is an abbreviation for two-phase Hájek.) Two other generalized regression estimators were considered. In both cases, x and z contains a variable with value one for all taxfilers. One generalized regression estimator involves calibration on taxfiler revenue during second-phase weighting. (Taxfiler revenue is included as a second auxiliary variable in z .) The second estimator involves calibration on taxfiler revenue at both phases of weighting. (Taxfiler revenue is included as a second auxiliary variable in both x and z .) Estimates of domain totals computed using these two estimators are denoted by $\hat{Y}_{\text{GREG-R2}}(d)$ and $\hat{Y}_{\text{GREG-R1R2}}(d)$, respectively, in subsequent discussion.

Estimates were produced for two variables of interest – transcribed revenue and total expenses. There are some conceptual differences between transcribed revenue and taxfiler revenue. For example, capital gains and extraordinary items are included in taxfiler revenue in many industries while they are excluded from transcribed revenue. In addition, taxfiler revenue contains more data capture errors than transcribed revenue since it is not subject to the same level of quality control.

The population used for the study included about 140,000 T2 taxfilers who reported over \$25,000 in revenue for tax year 1989. The first- and second-phase selection probabilities used during sampling for production for tax

year 1989 were employed. The first-phase sample included approximately 31,000 taxfilers and there were about 23,000 businesses in the second-phase sample. The correlation between taxfiler revenue and transcribed revenue for businesses in the second-phase sample was 0.969, while the correlation between taxfiler revenue and total expenses was 0.960.

Large proportions of units in the first- and second-phase samples were selected with certainty. All units with first-phase selection probability one were excluded from first-phase weighting and the corresponding g -weights were set to one. Units with second-phase selection probability one were treated analogously during second-phase weighting. There were 9,884 units in the first-phase sample with first-phase selection probabilities different from one and 910 units in the second-phase sample with second-phase selection probabilities different from one. Each first-phase poststratum consisted of one or more of the first-phase sampling strata used during sampling for 1989 production. These strata were defined using five revenue classes. All the sampling strata included in any particular first-phase poststratum corresponded to the same revenue class. Each first-phase poststratum contained a minimum of twenty sampled units. The use of a minimum sample size was motivated by concerns about the bias in $\hat{V}(\hat{Y}_{\text{GREG}}(d))$ when the number of sampled units used for estimation of regression coefficients is very small (Rao 1968b). If a first-phase sampling stratum included fewer than twenty sampled units, it was combined with sampling strata for similar SIC2 codes and the same revenue class until a poststratum containing at least twenty sampled units was obtained. Application of this procedure led to 166 first-phase poststrata. Second-phase poststrata were formed analogously, combining sampling strata for similar SIC4 codes to obtain a minimum sample size of twenty for each poststratum. There were 30 second-phase poststrata.

First and second-phase weights for $\hat{Y}_{\text{GREG-TPH}}(d)$, $\hat{Y}_{\text{GREG-R2}}(d)$ and $\hat{Y}_{\text{GREG-R1R2}}(d)$ were calculated using a modified version of the SAS macro CALMAR (Sautory 1991). The set of first-phase sampling weights calculated for the GREG-R1R2 estimator included twelve negative weights. There were no negative second-phase weights calculated for either GREG-R2 or GREG-R1R2. (Negative weights are not possible for the GREG-TPH estimator.) Estimates of transcribed revenue and total expenses were produced for 77 SIC2 domains, 256 SIC3 domains and 587 SIC4 domains using the three GREG estimators, as well as $\hat{Y}_{H-T}(d)$. Since GREG-R1R2 did not produce any negative estimates, no measures were taken to modify the negative weights associated with the estimator.

Results of comparisons of the GREG-TPH and H-T estimators are presented in Table 1 and Table 2. The mean gains and mean losses reported in the tables are averages of ratios of coefficients of variation. The GREG-TPH estimator performs better than the H-T estimator for the

Table 1

Comparison of GREG-TPH and H-T Estimators for Transcribed Revenue, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-TPH		Losses Using GREG-TPH	
	Number	Mean	Number	Mean
SIC2	57	0.768	20	1.113
SIC3	175	0.909	81	1.082
SIC4	359	0.945	228	1.079

Table 2

Comparison of GREG-TPH and H-T Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-TPH		Losses Using GREG-TPH	
	Number	Mean	Number	Mean
SIC2	57	0.773	20	1.100
SIC3	175	0.910	81	1.082
SIC4	355	0.945	232	1.079

majority of domains. The gains obtained using GREG-TPH are particularly large for SIC2 domains. At the SIC4 level, the estimated coefficient of variation (CV) for the GREG-TPH estimate of total expenses is lower than the estimated CV for the H-T estimate for 60.5% of domains. In cases in which the estimated CV for GREG-TPH is lower it is 5.5% smaller, on average, than the estimated CV for H-T. When the estimated CV for GREG-TPH is higher it is 7.9% larger than the estimated CV for H-T, on average. In addition to the information in Tables 1 and 2, there is another reason to prefer GREG-TPH to H-T. Each year, tax return information for some sampled taxfilers is not received by Statistics Canada or is unusable because it does not include the necessary financial statements. Assuming that such cases of nonresponse are ignorable, the GREG-TPH estimator provides an automatic nonresponse adjustment.

The results in Tables 1 and 2 indicate that the relative performance of the GREG-TPH and H-T estimators are very similar for both variables of interest. The results of the other comparisons of estimators done as part of this empirical study did not depend on the variable of interest in any important way. Consequently, only results for total expenses are reported in subsequent tables.

The GREG-TPH estimator is compared to GREG-R2 and GREG-R1R2 in Tables 3 and 4. Based on estimated coefficients of variation, GREG-R2 performs slightly better than GREG-TPH. Since a large proportion of units in the second-phase tax sample have second-phase selection probability one and both GREG-R2 and

GREG-TPH use the same auxiliary variables during first-phase weighting, the marginal differences between GREG-R2 and GREG-TPH are not surprising. Estimated CVs for GREG-R1R2 are generally smaller than estimated CVs for GREG-TPH and the relative performance of GREG-R1R2 improves as domain size increases. Nevertheless, GREG-R1R2 is superior to GREG-TPH for only 64% of SIC4 domains, and the average increase in estimated CVs for those domains in which GREG-R1R2 did worse than GREG-TPH is larger than the average decrease in estimated CVs for domains in which GREG-R1R2 performed better.

Table 3

Comparison of GREG-R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-R2		No Difference	Losses Using GREG-R2	
	Number	Mean	Number	Number	Mean
SIC2	38	0.993	26	13	1.001
SIC3	58	0.991	158	40	1.002
SIC4	88	0.988	439	60	1.009

Table 4

Comparison of GREG-R1R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation

Type of Domain	Gains Using GREG-R1R2		Losses Using GREG-R1R2	
	Number	Mean	Number	Mean
SIC2	51	0.867	26	1.170
SIC3	160	0.934	96	1.093
SIC4	377	0.954	210	1.074

The results in Tables 3 and 4 indicate that, although the GREG-R1R2 estimator shows some promise, it would be inappropriate to completely replace the GREG-TPH estimator currently used in production by GREG-R1R2. The improvements obtained using GREG-R1R2 are relatively marginal, given the strong correlation between taxfiler revenue and total expenses. Larger improvements could be obtained if: (i) SIC codes used for first-and second-phase stratification were always consistent with SIC codes used to determine the domain membership of sampled units; and (ii) formation of first-and second-phase poststrata did not require combination of sampling strata to obtain a minimum sample size in each poststratum.

The results reported in Table 5 were obtained after SIC codes assigned to taxfilers by Revenue Canada and SIC codes used for stratification of the second-phase sample were changed for sampled units, where necessary, to eliminate inconsistencies between these codes and those

Table 5

Comparison of GREG-R1R2 and GREG-TPH Estimators for Total Expenses, Estimated Coefficients of Variation, No Misclassification

Type of Domain	Gains Using GREG-R1R2		Losses Using GREG-R1R2	
	Number	Mean	Number	Mean
SIC2	66	0.778	11	1.057
SIC3	184	0.916	72	1.047
SIC4	402	0.944	185	1.034

used to determine domain membership. A comparison of Tables 4 and 5 indicates that the relative performance of GREG-R1R2 is considerably better when there are no classification errors. GREG-R1R2 reduces estimated CVs by over 22% (on average) for over 85% of SIC2 domains.

Throughout the empirical results reported here, performance improvements obtained through the use of additional auxiliary information increase as domain size increases. This result is consistent with the observations in Section 2 concerning the conditions under which correlations between $y(d)$ and the vectors of auxiliary variables, x and z , will be high. Provided that the variable of interest and the auxiliary variables are highly correlated, correlations involving $y(d)$ will be strong if each poststratum containing at least one sampled unit falling in domain d also contains relatively few sampled units that do not fall in domain d .

5. CONCLUSIONS

Generalized regression estimation provides a convenient framework for the use of auxiliary information. A generalized regression estimator for a two-phase sample design with Poisson sampling at both phases of selection is derived in this paper. The efficiency of the estimator is investigated through application to the two-phase tax sample selected by Statistics Canada to obtain annual estimates of the economic activity of small businesses. The estimation method currently used in production for this survey incorporates poststratified ratio adjustments during both first-and second-phase weighting to compensate for differences between actual and expected sample sizes. This poststratified estimator is a particular case of the generalized regression estimator.

In an empirical study, the generalized regression estimator currently used in production (GREG-TPH) performs much better than the Horvitz-Thompson estimator. Two other generalized regression estimators are also compared to GREG-TPH. The alternative estimators produce improvements for large domains. However, their performance for the smaller domains that are of particular interest to users

of estimates based on the two-phase tax sample does not justify complete replacement of the current production methodology.

ACKNOWLEDGEMENTS

The authors would like to thank René Boyer for providing a modified version of the SAS macro CALMAR suitable for the empirical study, as well as K.P. Srinath and Michael Hidioglou for helpful discussions. Thanks are also due to Michael Bankier and Jean Leduc for helpful comments on a earlier draft of this paper.

APPENDIX A: DERIVATION OF VARIANCE OF $\hat{Y}_{\text{GREG}}(d)$ AND VARIANCE ESTIMATOR

The variance of $\hat{Y}_{\text{GREG}}(d)$ can be derived using the identity

$$V(\hat{Y}_{\text{GREG}}(d)) = E_1 V_2(\hat{Y}_{\text{GREG}}(d)) + V_1 E_2(\hat{Y}_{\text{GREG}}(d)).$$

First, consider the variance of the estimator with respect to the second phase of sampling, conditional on the results of first-phase calibration. If the vector of auxiliary variables for second-phase weighting, z , includes a variable with value one for all taxfilers (or a linear combination of auxiliary variables that is equal to one for all taxfilers can be constructed), the generalized regression estimator can be written as

$$\begin{aligned} \hat{Y}_{\text{GREG}}(d) &= \sum_{i \in s_2} w_{1i} w_{2i} y_i(d) \\ &= \sum_v \sum_{i \in s_2 \cap v} w_{1i} (y_i(d) - z_i' \hat{B}_v) / p_{2i} + \sum_v \tilde{Z}_v \hat{B}_v. \end{aligned}$$

Ignoring the variability due to the estimation of regression coefficients during second-phase weighting, we have

$$\begin{aligned} E_1 V_2(\hat{Y}_{\text{GREG}}) &\approx E_1 V_2 \left(\sum_{i \in s_2} w_{1i} Q_{2i} / p_{2i} \right) \\ &= E_1 \left(\sum_{i \in s_1} \frac{(1 - p_{2i})}{p_{2i}} w_{1i}^2 Q_{2i}^2 \right). \end{aligned}$$

The estimator of $E_1 V_2(\hat{Y}_{\text{GREG}}(d))$ based on the variance estimator for calibration estimators advocated by Deville and Särndal (1992, p. 380) is

$$\hat{S}_1 = \sum_{i \in s_2} \frac{(1 - p_{2i})}{(p_{1i} p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Ignoring variability due to the estimation of regression coefficients during first-phase weighting, the second term in the variance expression can be written as

$$\begin{aligned} V_1 E_2(\hat{Y}_{\text{GREG}}(d)) &= V_1 \left(\sum_{i \in s1} w_{1i} y_i(d) \right) \\ &= \sum_i \frac{(1 - p_{1i})}{p_{1i}} Q_{1i}^2. \end{aligned}$$

An estimator of this term is

$$\hat{S}_2 = \sum_{i \in s2} \frac{(1 - p_{1i})}{p_{1i}^2 p_{2i}} (g_{1i} q_{1i})^2.$$

REFERENCES

- ARMSTRONG, J., BLOCK, C., and SRINATH, K.P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, 11, 407-416.
- BANKIER, M., RATHWELL, S., and MAJKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census of Population. *Statistics Sweden, Workshop on the Uses of Auxiliary Information in Surveys*.
- BREWER, K.R.W., EARLY, L.J., and JOYCE, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.
- CHOUDHRY, G.H., LAVALLÉE, P., and HIDIROGLOU, M. (1989). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.
- DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 34, 911-934.
- DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- HIDIROGLOU, M.A., SÄRNDAL, C.-E., and BINDER, D.A. (1993). Weighting and estimation in establishment surveys. Paper presented at the International Conference on Establishment Surveys, Buffalo, New York.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- RAO, J.N.K. (1968a). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.
- RAO, J.N.K. (1968b). Some small sample results in ratio and regression estimation. *Journal of the Indian Statistical Association*, 6, 160-168.
- ROYALL, R.M., and EBERHARDT, K.R. (1975). Variance estimates for the ratio estimator. *Sankhyā, Series C*, 37, 43-52.
- SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAUTORY, O. (1991). La macro SAS: CALMAR. Unpublished manuscript, Institut national de la statistique et des études économiques, Paris.
- STATISTICS CANADA (1980). *Standard Industrial Classification*. Catalogue No. 12-501E, Statistics Canada.
- SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: A useful technique. *Journal of Official Statistics*, 2, 161-168.
- WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.
- WU, C.F.J., and DENG, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In Box, G.E.P. *et al.* (Eds.), *Scientific Inference, Data Analysis and Robustness*, New York: Academic Press, 245-277.

Two-Stage Area Frame Sampling on Square Segments for Farm Surveys

F.J. GALLEGO, J. DELINCÉ and E. CARFAGNA¹

ABSTRACT

In the MARS Project (Monitoring Agriculture with Remote Sensing) of the E.C. (European Community), area frames based on a square grid are used for area estimation through ground surveys and high resolution satellite images. These satellite images are useful, though expensive, for area estimation: their use for yield estimation is not yet operational. To fill this gap the sample elements (segments) of the area survey are used as well for sampling farms with a template of points overlaid on the segment. Most often we use a fixed number of points per segment. Farmers are asked to provide global data for the farm, and estimates are computed with a Horvitz-Thompson approach. Major problems include locating farmers and checking for misunderstanding of instructions. Good results are obtained for area and for production of the main crops. Area frames need to be complemented with list frames (multiple frames) to give reliable estimates for livestock.

KEY WORDS: Area frame; Point sampling; Segment sampling; Farm sampling.

1. INTRODUCTION

The main purpose of this paper is to present the method used to sample farms in an area frame by the MARS (Monitoring Agriculture with Remote Sensing) Project of the European Community (EC). Sampling farms is not a central activity in this project, but rather a way of bypassing the limitations of the actual capacity of satellite images, especially for yield estimation. We shall present a brief overview of the MARS Project to make up for the few existing references in statistical journals (Ambrosio 1993, Gallego 1992). Other presentations can be found in conference papers (Meyer Roux 1990, Delincé 1990, Sharman *et al.* 1992, Carfagna *et al.* 1994) or remote sensing journals (González *et al.* 1991, Gallego *et al.* 1993).

2. THE MARS PROJECT OF THE EUROPEAN COMMUNITY

The MARS Project was launched in 1988 to assess and to develop operational applications of Remote Sensing to Agricultural Statistics. It is carried out by the Institute of Remote Sensing Applications (IRSA) of the Joint Research Centre (JRC) of the EC. Most of the activities of the period 1988-1993 were divided into 4 main parts, named "actions":

- (1) Regional Crop Inventories.
- (2) Monitoring Vegetation.
- (3) Agrometeorological Models.
- (4) Rapid Estimates at the EC level.

Some work is made as well in other related fields, such as area frame sampling. We shall focus here on a sampling

method used in the frame of action 1 "Regional Inventories", but we shall first say a word about the other actions.

2.1 Monitoring Vegetation

This action deals with low resolution satellite images from NOAA-AVHRR (Advanced Very High Resolution Radiometer). In these images each pixel has about 1 km² in the vertical of the satellite orbit. The main objectives are the development of friendly software for the pre-treatment of these images, and building a data bank with time series vegetation indexes and other indicators for about 3,000 monitoring units in the EC. These monitoring units have not yet been definitely defined. They should be geographic areas roughly between 500 km² and 1,000 km² with a more or less homogeneous vegetation or greenness index (Houston 1984, Goward 1991).

2.2 Agrometeorological Models

General and crop specific models are being currently developed on the basis of data from a network of about 650 Meteorological Observatories in Europe and surrounding areas. This model CGMS (Crop Growth Monitoring System), developed in collaboration with the WOFOST (World Food Studies Centre, in Wageningen, Netherlands), also uses other data, such as soil and elevation data, together with information on the physiology of plants (van Diepen 1989, van Lanen 1992). Remote sensing (low resolution images) will come into the picture later for the spatial interpolation of ground observed meteorological data. Parameters of the model are currently computed for each cell of a 50 km × 50 km grid.

¹ F.J. Gallego, Joint Research Centre of the European Communities, tp. 440, 21020 Ispra, Varese, Italy; J. Delincé, Commission of the E.C. DG VI, Loi 120, 4-23, 1049 Brussels, Belgium; E. Carfagna, Department of Statistics, University of Bologna, V. Belle Arti 41, 40126 Bologna, Italy.

2.3 Rapid Estimates at the E.C. Level

The main goal is giving rapid estimates of area and yield change of annual crops compared with the previous year based on a two-stage sampling scheme: 53 sites (Figure 1) of $40 \text{ km} \times 40 \text{ km}$ with a sample of 16 squared segments of $700 \text{ m} \times 700 \text{ m}$ (Figure 2) in each of the sites. Individual data are acquired by photo-interpretation of SPOT-XS or Landsat-TM images. An average of three images is analysed for each site with a minimum of ground information, namely a general knowledge of the dominant crops in each area. A ground survey is made for an a posteriori validation of the photo-interpretation. A monthly report (from March to November) is produced with an update of the estimates. Each report should use all the images acquired more than 15 days before.



Figure 1. Sample of 53 sites for rapid crop estimates in the E.C.

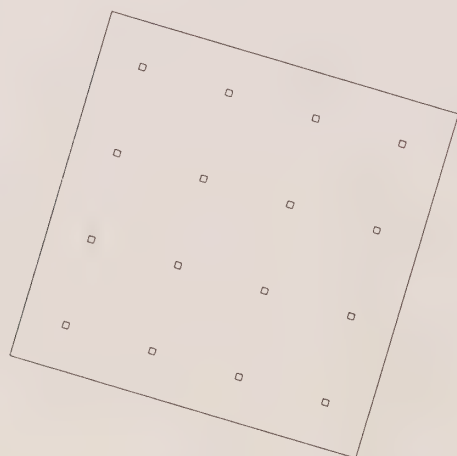


Figure 2. Segments in one site (rapid estimates in the E.C.)

2.4 Regional Crop Inventories by Segment Sampling and Remote Sensing

The objective of the action was to implement, to adapt and to assess estimation methods for crop area and production based on area frame sampling and satellite images. When this action was implemented by the IRSA in 1988 on five pilot regions of approximately $20,000 \text{ km}^2$ each; an absolute priority was given to annual crops: soft and durum wheat, barley, rapeseed, dried pulses, sunflower, maize, cotton, tobacco, sugar beet, potatoes, rice and soya, as well as fallow. Attention is being shared more and more by permanent crops, pastures, and non-agricultural land uses.

Since 1990 the IRSA has progressively transferred the initiative to regional or national administrations that wish to use area frame surveys based on segments. In general, the activities have been shifted to the southern countries of the EC and the former communist countries in central Europe, that have shown much interest in the method (Figure 3). In some cases, like in Italy, there is just an exchange of points of view between the national project and the IRSA.

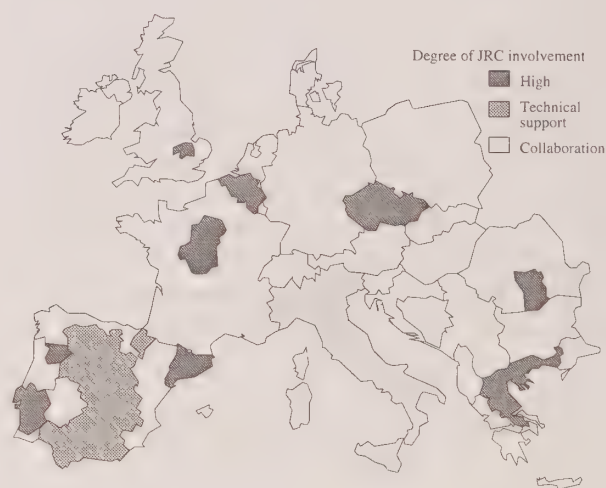


Figure 3. European regions with segment surveys in 1992.

2.4.1 Sampling Segments on a Square Grid

There are two main approaches to building an area frame based on segments: the segments can be drawn on topographic or cadastral maps following roads, rivers, or limits of fields (sometimes called cadastral segments). The sample is usually drawn with a two-stage procedure with intermediate primary sample units to reduce the burden to build the frame (Cotter 1987), which remains in any case a heavy operation.

We generally use frames based on a square grid (Gallego and Delincé 1994), which is much faster to define. Satellite images are generally (but not necessarily) used for stratification prior to sampling.

Figure 4 illustrates a small example of this kind of sample with a very simple stratification and segments of 25 ha (hectares). Sampling is systematic, repeating a pattern in square blocks. In this case the blocks have a size of 10 km \times 10 km, and the pattern has 4 replicas in the most agricultural stratum (plain), 2 replicas in the hills, and one in the mountains.

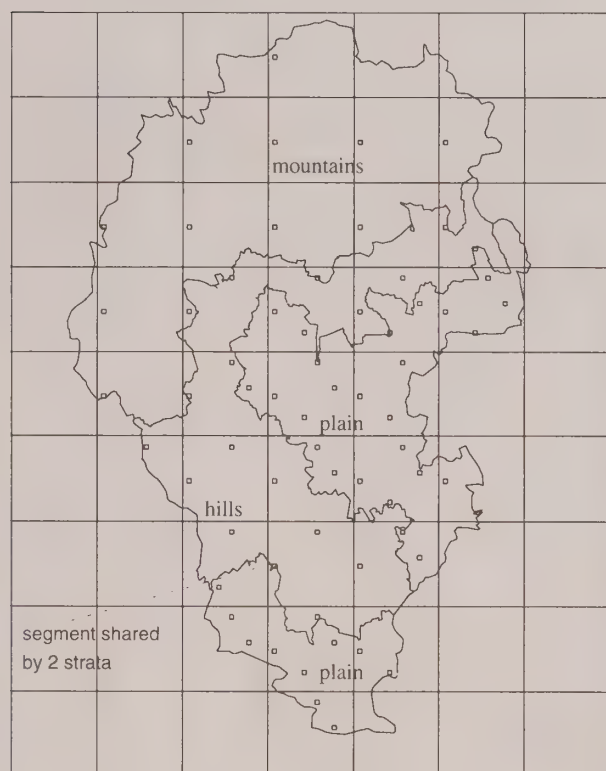


Figure 4. Example of area frame sample with squared segments and squared blocks.

The main drawback of this approach is the management of segments that fall on the boundary between two strata (Figure 5). Three alternatives are being tested for this problem: (1) adapting the stratification to the sampling grid, (2) splitting border segments into pieces that belong to different strata, and (3) keeping only the largest one among these pieces.

The most frequent non-sampling errors – shifts in location and inaccuracy in shape or size of the segment – are not strongly correlated with land use. No major influence has been found on the area estimates or their precision.

The sample pattern to be repeated in each block is drawn at random with a restriction on the distance between segments in order to avoid segments that are too close to

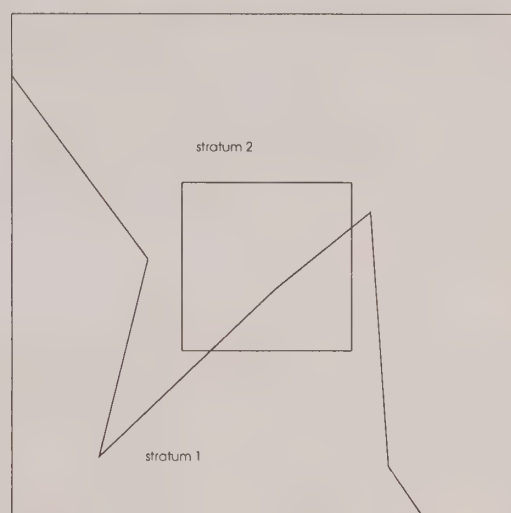
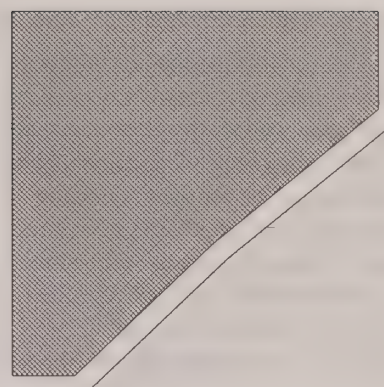


Figure 5. A segment can be split by a stratum boundary.

each other. Cluster estimators can be used in this case rather than standard formulae for random sampling (Fuentes 1994, Ambrosio 1993). Systematic sampling has a risk of bias if there is a cyclic effect in the landscape with a period that coincides with the block size (10 km in the example), but this is very unlikely. The distance threshold between segments can induce an overestimation of standard errors if the spatial correlation is significantly positive for distances less than the threshold.

The size of the segments varies from region to region depending on the agricultural landscape, especially on the size of fields. In the Czech Republic, the segment size was 400 ha. For the area survey, enumerators locate the segments, draw fields on a transparent sheet placed over an aerial photograph, and write down their land use. About 5% to 10% of the segments are visited again by supervisors to check for possible errors on the ground work. Satellite images are not used either for the survey itself or for the farm survey, but they can be optionally used to improve the precision of the area estimates as described in the next section.

2.4.2 Improving Area Estimates with Satellite Images

High resolution satellite images from Landsat-TM or SPOT-XS sensors have been assessed and are still being used at moderate scale to improve the estimates obtained from the ground survey on a sample of segments. The most commonly used approach is the regression estimator on classified images. An alternative estimator based on confusion matrices has been tested with results that are very close to those of the regression estimator (Hay 1988, Gallego 1994).

The conclusions of this assessment are similar to those of the US Department of Agriculture (Allen 1990). The use of satellite images for area estimation is operational, but still too expensive for the efficiency obtained. The economic threshold can be reached by improving image processing automation, since the cost of image processing in the European market for this purpose is much higher than the cost of the images themselves. This threshold has nearly been reached with Landsat-TM images in Greece. Different conclusions on cost analysis are presented by Giovacchini (1992).

3. SAMPLING FARMS BY POINTS

For agricultural surveys in the European Community, farms are traditionally sampled from a list frame (Eurostat 1991). The list is a census of farms that exceed a certain size threshold. In many countries an agricultural census is made every 10 years and is seldom updated (if ever). Hence there may be a substantial difference between the sampling frame and the actual population at the date of the survey. The situation is worse in the central European countries of the former Eastern Block (the area between Poland and Rumania-Bulgaria), where the change of land property structure is so rapid that the census may not exist for private farms and becomes obsolete for co-operatives.

Area frames on square segments can be easily defined when the geographic borders of the region are known. A subsample of these segments is used as well for sampling farms in several countries with the help of a template of points overlaid on the segment. This has been experimentally tested in Germany, Portugal, Italy (Carfagna 1991) and Spain, and is now being regularly used in Greece, Rumania and the Czech Republic.

The template is the same for all the segments in a stratum, and usually symmetric to reduce the risk of bias due to a particular geographic location. Data are obtained only for farms corresponding to points falling on Utilized Agricultural Area (UAA).

The definition of UAA used in the field work is adapted to each national system. Farm buildings and rough pastures are included in some countries and excluded in other countries. The crucial point is that the definition used must be consistent with the definition of the column UAA used for computation (Table 1).

Table 1

Observations Generated by Points Sampled in the Segment of Figure 6

Segment	Point	UAA	Perma- nent Crops	Wheat		Barley	
				Area	Produc- tion	Area	Produc- tion
1	1	19	4	12	64	0	0
1	2	0	0	0	0	0	0
1	3	0	0	0	0	0	0
1	4	35	0	24	131	3	12
1	5	35	0	24	131	3	12
2

In the example of figure 6, point 3 fell on woodland and point 2 on a built area. They will generate two zero-valued records in the farm file. The enumerator will have to locate the farmers for the other three points. The farm corresponding to point 1 has other fields in the segment, that will be implicitly included in the survey, but the enumerator will not need to find out if these fields exist. Points 4 and 5 belong to the same farm, and it will appear twice in the farm file (Table 1).



Figure 6. Segment with a pattern of 5 points for farm sampling.

Farmers are located and asked to provide global data for the farm, including total area and production of each target crop. No question is asked about the production of each field or the set of fields inside the segment. This is not necessary because in the final formulae to compute the estimates (formulae 2 and 3 in section 4.1) the crop area or the production in the tract is not used.

The ground survey instructions are usually transferred from the JRC to National Administrations. They explain the instructions to Regional co-ordinators, who give the information to the enumerators. Instructions may be modified in some of these steps. Checking that the instructions have not been misunderstood is sometimes difficult, in part because linguistic limitations are a serious barrier to direct contact with enumerators. In some countries (e.g., Spain) farmers live mainly in rather large urban nuclei and are difficult to locate; this can lead to a significant amount of missing data.

4. ESTIMATES BASED ON FARMS SAMPLED BY POINTS

We assume that the population Ω of segments is divided into strata Ω_h , $h = 1, \dots, H$, the total population size is N segments (N_h for stratum Ω_h) and the sample size is n segments (n_h). The size of our sample of points in each segment will be K_i , previously fixed; in general we have $K_i = K$, constant across all strata, out of which F_i correspond to the farms on which these points fall. Each segment i has a total UAA surface U_i .

We have a two-staged sampling scheme. In the first stage the segment i is selected with probability $p_i = 1/N_h$ in each of the n_h trials. In the second stage the unit is not the farm but the tract (UAA in a segment, that belongs to the same farm). The tract k of segment i has an area T_{ik} . The total UAA of the farm is A_{ik} over all segments. U_i is the sum of the tracts T_{ik} in segment i .

The method presented below is closely related to the so called "weighted segment estimator" approach used in the U.S. and in Canada (Nealon 1984).

4.1 Estimates Based on Farms and Non-Farm Points

There will be $K - F_i$ observations (fictitious farms) with value 0 corresponding to points outside the UAA.

Sampling through points means that tracts are selected with replacement and with a probability p_{ik} proportional to the area T_{ik}/D_i , (the knowledge of T_{ik} is not necessary), where D_i is the size of the segment determined by the frame design. We are implicitly assuming that the surveyed region is flat. A slight bias might be introduced by the fact that annual crops are usually on more or less flat land and pastures or non-UAA are often on land with a steeper slope.

The sampling is done with replacement: a farm can be selected more than once, which gives easier formulae for variance estimation. Strictly speaking the joint selection probability that farms k and k' are in the sample $p_{ikk'} \neq p_{ikk} \times p_{ikk'}$ as would be the case if the different points of the template were drawn independently, since there is usually a relatively large distance between them. We will disregard this fact in this paper.

W_{ik} will be an additive quantity for a farm, most often the production or the area of a particular crop. It is obvious that yield is not an additive variable.

Since we have no information about how W_{ik} is distributed inside the farm, we create a fictitious variable X that is uniformly distributed, and that has, by definition, the same total as W for each farm:

$$X_{ik} = \frac{T_{ik}}{A_{ik}} W_{ik}. \quad (1)$$

Estimating the totals of X and W are equivalent problems.

The two-stage version of the Horvitz-Thompson estimator for the total of X in the stratum Ω_h gives:

$$\begin{aligned} \hat{X}_h &= \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{X}_i}{p_i} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{K_i} \sum_{k=1}^{K_i} \frac{X_{ik}}{p_{ik}} = \\ &= \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{D_i}{K_i} \sum_{k=1}^{K_i} \frac{W_{ik}}{A_{ik}}. \end{aligned} \quad (2)$$

This means that, even if the second stage sampling unit is the tract, we do not need to know its area nor X_{ik} , but just the global information about the farm.

The estimator is a linear function of the estimates on the selected segments. Its variance in stratum Ω_h can be estimated as (Cochran 1977, section 11.6):

$$\begin{aligned} \hat{V}(\hat{X}_h) &= \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \sum_{i=1}^{n_h} \frac{(\hat{X}_i - \bar{X}_h)^2}{n_h - 1} + \\ &= \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{K_i(K_i - 1)} \sum_{k=1}^{K_i} \left(\frac{W_{ik} D_i}{A_{ik}} - \hat{X}_i \right)^2. \end{aligned} \quad (3)$$

The estimates for the total are:

$$\hat{X} = \sum_{h=1}^H \hat{X}_h \quad \hat{V}(\hat{X}) = \sum_{h=1}^H \hat{V}(\hat{X}_h). \quad (4)$$

Crop areas are currently estimated from the segment survey with more objective ground data (direct observation of the enumerator on the field), although some bias can appear due to the imperfect location of the segments on the ground. Farm surveys provide both area and production estimates, but they can have more significant bias due to non response and to a subjective tendency of the farmer that can depend on whether he is more concerned about taxes or about subsidies at the time of the survey. Comparing both area estimates, from segment survey and farm survey, can be useful to check for possible bias on the production estimate based on the farm survey.

The estimates are also possible for cattle, but the results will be presumably bad if there are a substantial number of farms without any UAA, which will not be sampled: the coverage of the area frame will not be complete in this case. On the other hand it may happen that the number of livestock does not correlate to the UAA and hence to the probability of selection. This results in inefficient estimates.

A program in C for Personal Computers has been written (Dicorato 1993) to compute estimates using this method. The main part of the program was first written to compute estimates on a segment survey.

4.2 Estimation Based Only on Farm Points

We shall mention another option that consists of using only points that fall in the UAA. In this case, we first fix F_i , the number of points that fall in UAA (often $F_i = F_h$, constant in each stratum). In segment i we observe as many points as necessary to have F_i points in the UAA. If the segment i has no UAA, one observation (fictitious farm) is added with 0 values. This is actually an implicit second-stage stratification or stratification of the first-stage units (segments) into two strata; UAA and non-UAA. The non-UAA stratum is not sampled. In this case (2) and (3) are to be adapted substituting K_i by F_i and D_i by U_i . Some inconsistency may arise in hilly areas because A_{ik} comes from the farmer's declaration and U_i from segments drawn on the ground over aerial photographs.

$$\hat{X}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{X}_i}{p_i} = \frac{N_h}{n_h} \sum_{h=1}^{n_h} \frac{1}{F_i} \sum_{k=1}^{K_i} U_i \frac{W_{ik}}{A_{ik}}, \quad (5)$$

$$\hat{V}(\hat{X}_h) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \sum_{i=1}^{n_h} \frac{(\hat{X}_i - \bar{X}_h)^2}{n_h - 1} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{F_i(F_i - 1)} \sum_{k=1}^{K_i} \left(\frac{W_{ik} U_i}{A_{ik}} - \hat{X}_i\right)^2, \quad (6)$$

the second term of (6) is null for segments with no UAA. This term cannot be computed if $F_i = 1$ because of non-response. A value 0 can be attributed, though this will lead to an underestimation of the within-segment variance, which is relatively small according to calculations made on available data (Carfagna 1992).

This approach has only been used once to resolve a misunderstanding of instructions for ground work that should have been performed following the method in section 4.1. However advantages and drawbacks of both approaches are not clear, and no systematic comparison has been made so far on the same region and same year. Using only farm points can increase the cost of the survey if the number of points per segment is to be kept constant,

but the non-UAA points removed correspond to null values of W_{ik} , and their removal can result in a reduction of the variance.

4.3 Farms with Fields in Different Strata

At first sight, the estimator (2) seems to assume that a farm k that has been selected through a point in stratum Ω_h is completely included in this stratum. It is obvious that a farm can have fields in different strata, and the question arises as to whether this fact disturbs the reliability of the results.

We stress again that the variable used is not really W_{ik} , but X_{ik} defined for each individual tract. The total of W does not coincide with the total of X in each stratum, but it does in the whole region as long as

$$\sum_i T_{ik} = A_k. \quad (7)$$

Notice that A_k is identical to what we have called previously A_{ik} , where the subindex i is used only to indicate that farm k has been selected in the sample through segment i .

This identity holds on the population, regardless of the sampling procedure, if the farms are entirely inside the region and if the geometry of the ground survey document (aerial photograph) is correct.

The perturbation due to farms with fields in different regions is expected to be small because of the low proportion (generally under 1–2%) and because there is a compensation between the bias due to fields inside the region belonging to farms with the headquarters outside the region and vice versa. We are assuming that the total of W is calculated on the farms that have their headquarters inside the surveyed region.

4.4 Nonresponse

We refer here to the estimators based on farm and non-farm points (section 4.1). If a farmer does not co-operate or cannot be found, the corresponding row or rows of the input table (Table 1) are substituted with the average values of responding farms in the segment, if there are any; otherwise they are substituted with the average of responding farms for all the segments in the current stratum.

If in the second stage (sampling farms inside the segment) we consider farm and non-farm points, and give value 0 to the points that fall in non agricultural land, it is obvious that the exclusion of nonrespondents would produce a serious bias, because the zero values corresponding to non-UAA are never missing. These points are not used to compute the “average farm” values used to fill missing values. There is still a risk of bias if farmers who cannot be located or refuse to co-operate have a peculiar behaviour, e.g., if they are on the average smaller or less efficient farms.

We could have considered a different way of overcoming this problem: eliminating both missing values and a proportional number of 0 values corresponding to non-UAA points. Both give the same estimate for the total, but the second solution is more uncomfortable because the sample size in the second stage is not an integer any more.

The introduction of "average farm" values will lead to a negative bias on the variance. To compensate it, the farm is not included in the sample size K_i for the computation of variances.

5. RESULTS: TWO EXAMPLES

We discuss below some results from two regions: Emilia Romagna (Italy) and the Czech Republic. In the Czech Republic, the method presented in sections 2.4, 3 and 4.1 was used; there were no missing data at all. In Emilia Romagna the general design of the survey did not follow exactly the procedure outlined above. Missing data were treated as stated in section 4.4.

5.1 Emilia Romagna 1990

In Emilia Romagna an area of 19,500 km² was divided into 4 strata, excluding mountainous areas. A sample of 313 "cadastral" segments (with physical boundaries) was drawn based on a two-staged procedure with primary sampling units (psu) of about 10 km². Segment size was approximately 50 hectares or 100 hectares, depending on the strata. 5 points per segment were drawn at random from a grid with a 50 metre step.

Out of the 1,565 points sampled: 326 were non-UAA, the farmer's address could not be found for 206 UAA points, 38 farmers were not located and 32 refused to co-operate. 963 UAA points from 285 segments had valid data, corresponding to 617 farms, some of which appear more than once in the sample.

When we think only of area estimation, the segment survey can be seen as more objective and complete, since there are no missing data and observations do not rely on farmers' answers. If we accept this principle we can have an idea of a possible bias in the farm survey by comparing with the area estimates of the segment survey. Estimates can be compared in Table 2 for the main crops in the region. Figures match well for cereals, excepting durum wheat, and permanent crops, but some problems appear for sugar beet and soya, that might be related to misunderstandings on how to declare second crops in the same year and the same field, or with a bias due to missing values. Official statistics are produced taking into account a variety of information. Durum wheat is reported separately because of the special meaning of this crop due to the significant subsidy granted by the EC to each hectare of crop.

Table 2

Results of the Segment Survey and the Farm Survey for Main Crops in Emilia Romagna (1990)

Emilia Romagna	Segment Survey		Farm Survey				ISTAT
	Area × 1,000 ha		Area × 1,000 ha		Prod. × 1,000 tm		Area
	Esti- mation	CV %	Esti- mation	CV %	Esti- mation	CV %	
Soft wheat	212	5.7	208	6.9	1,177	8	212
Durum wheat	46	14.9	48	15.2	260	14	72
Barley	43	11.2	50	17.7	184	17	38
Rice	—	—	4	59.0	23	61	6
Sugar beet	111	7.1*	96	9.6	5,474	28	119
Soybeans	76	6.0*	55	11.6	321	39	47
Vineyards	78	13.3*	76	18.7			75
Orchards	91	13.1*	96	19.7			85

* Estimate corrected by regression on classified satellite image.

ISTAT: Official statistics. No precision provided.

The coefficients of variation in the farm survey have a reasonable behaviour for cereals, but become more difficult to understand for sugar beet and soybeans. The high CV (Coefficient of variation) for the production can be due to higher yields in larger, more specialized farms.

A correction of the production estimate can be made using the difference of area estimates between the segment survey and the farm survey. A regression estimator approach might be a good solution.

Livestock is seriously underestimated (Table 3) since many livestock owners do not have agricultural land. A mixed approach was used for cattle and pigs with an exhaustive survey using a list frame for the 50 largest farms and point sampling for the rest. The procedure works for pigs, but CVs are not yet satisfactory.

Table 3

Results of the Farm Survey on Area Frame and Mixed Frame for Livestock in Emilia Romagna (1990)

× 1,000 Units	Census	Area Frame		Mixed Frame	
		Estimate	CV %	Estimate	CV %
Cattle	869	829	14	894	13
Pigs	1,876	1,312	37	1,818	27
Sheep	90	38	74		

5.2 Czech Republic 1992

Area frames seem especially useful in the former communist countries in Europe because of the rapid change of property structure. Agricultural statistics are mainly produced with no sampling error by adding the data reported by each state farm or co-operative. This procedure will collapse in the coming years. It will be extremely

difficult to have an idea of the number of existing farms, and an agricultural census will be out of date before the data are elaborated. Area frames might be the best alternative.

The territory of the Czech Republic (about 80,000 km²) has been stratified into 6 strata by photo-interpretation of Landsat-TM images. The stratification needed 15 working days for one person. In 1992, a survey was made with a sample of 417 square segments of 400 ha drawn by repetition of a fixed pattern on blocks of 40 km × 40 km. Segments were visited and area estimates obtained as explained in section 2.4.1.

Farms have been sampled using a fixed grid of 5 points in each segment. The shape of the 5-point grid was in "x" like in figure 6. This procedure gave 2,085 points: 858 non-agricultural, and the other 1,227 from 458 farms. No missing data were recorded: all the farms were identified and none refused to co-operate. This happened mainly because the old structure of large farms was still nearly intact.

Table 4 compares the results of the segment survey (direct observations on the field), the farm survey (farms sampled by points), and official statistics for the main crops in the country. Official statistics are obtained by adding figures reported by all the state farms or co-operatives. There is a moderate disagreement on area estimates for wheat, maize, and potatoes. We should not exclude a bias in farmers' answers that has to do with self-consumption of agricultural products.

Table 4

Results of the Segment Survey and the Farm Survey in the Czech Republic (1992)

× 1,000 ha	Segment Survey		Farm Survey				CSO	
	Area	CV %	Area	CV %	Prod.	CV %	Area	Prod.
Wheat	824	5.4	757	3.7	3,412	4.9	780	3,413
Barley	655	5.1	630	3.8	2,521	4.3	640	2,512
Rapeseed	140	11.6	137	6.8	310	7.5	136	296
Sugar beet	119	11.5	127	8.1	4,172	11.0	125	3,874
Maize	361	7.5	326	4.8	8,884	4.3	361	8,904
Potatoes	109	13.6	92	7.9	1,706	8.7	111	1,969

CSO: Czech Statistical Office.

The coefficients of variation (CV) of the area estimates are lower in the farm survey than in the segment survey. This is not surprising since the farm survey gives information about fields outside the segments. The 458 selected farms represent more than 15% of the total UAA in the country. The CVs for production estimates are slightly higher than for area estimates (even lower in the case of maize). This seems to indicate that the variability of yields contributes less than the variability of areas to the variability of production.

6. CONCLUSIONS AND RECOMMENDATIONS

Area frames based on square grids are a pragmatic alternative to area frames based on ground elements delimited by physical features. They are much cheaper to build and they do not seem to have major drawbacks regarding the final results. However some theoretical work is still needed to determine under which conditions the location errors due to non-physical limits have a negligible effect on the estimates.

Sampling points inside area segments provides a feasible way to build frames for farm sampling. They are extremely useful if list frames (census) are poorly updated or do not exist. Sampling a few points per segment can be much cheaper than surveying all the farms with fields in the segment. Five points per segment seems to be a reasonable choice.

Area frames alone give poor results for livestock when the number of units is not strongly correlated with Utilized Agricultural Area of the farm.

ACKNOWLEDGEMENTS

We are grateful to the various National and Regional Administrations that have collaborated for this work, and to A. Burrill and O. O'Hanlon for the kind revision of the paper. The numerous comments by the referees have been essential for the paper to become more useful to readers.

REFERENCES

- ALLEN, J.D. (1990). A look at the remote sensing applications program of the national agricultural statistics service. *Journal of Official Statistics*, 6, 393-409.
- AMBROSIO, L., ALONSO, R., and VILLA, A. (1993). Estimación de superficies cultivadas por muestreo de áreas y teledetección. Precisión relativa. *Estadística Española*, 35, 91-103.
- CARFAGNA, E., RAGNI, P., ROSSI, L., and TERPESSI, C. (1991). Area frame: un Nuovo Istrumento per la Realizzazione delle Statistiche Agricole in Italia. *Contributi alla Statistica Spaziale*. University of Parma.
- CARFAGNA, E., and DELINCÉ, J. (1992). Farm survey based on area frame sampling. The case of Emilia Romagna in 1990. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publication of the E.C. Luxembourg.
- CARFAGNA, E., and GALLEGO, F.J. (1994). Extrapolating intra-cluster correlation to optimize the size of segments in an area frame. Conference on Applied Statistics to Agriculture, Kansas State University, Manhattan, KS.
- COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons.

- COTTER, J., and NEALON, J. (1987). Area frame design for agricultural surveys. U.S. Department of Agriculture. National Agricultural Statistics Service.
- DELINCÉ, J. (1990). Un premier bilan de l'action 1 Inventaires Régionaux du Projet Agriculture après deux années d'activité. Conference on the Application of Remote Sensing to Agricultural Statistics, Varese. Office for Publication of the E.C. Luxembourg.
- DICORATO, F. (1993). AIS estimation programs. User documentation. JRC Ispra.
- VAN DIEPEN, C.A., WOLF, J., VAN KEULEN, H., and RAPOLDT, C. (1989). WOFOST: A simulation model for crop production. *Soil Use and Management*, 5, 16-24.
- EUROSTAT 1991. Working party, Crop Products Statistics. Methodological reports. Document AGRI/PE/333, Luxembourg.
- FUENTES, M., and GALLEGO, F.J. (1994). Stratification and cluster estimator on an area frame by squared segments with an aligned sample. Conference on Applied Statistics to Agriculture, Kansas State University, Manhattan, KS.
- GALLEGO, F.J. (1992). Flächenschätzungen für einjährige Feldfrüchte mit Hilfe Fernerkundung. *Neue Wege raumbezogener Statistik. Forum der Bundesstatistik*, 20, 109-120. Wiesbaden: Statistisches Bundesamt.
- GALLEGO, F.J. (1994). Using a confusion matrix for area estimation with remote sensing. *Atti Convegno AIT*, Roma, 99-102.
- GALLEGO, F.J., and DELINCÉ, J. (1994). Area estimation by segment sampling. In *Euro-Courses Remote sensing applied to Agricultural Statistics*.
- GALLEGO, F.J., DELINCÉ, J., and RUEDA, C. (1993). Crop area estimates through remote sensing: Stability of the regression correction. *International Journal of Remote Sensing*, 14, 18, 3433-3445.
- GIOVACCHINI, A. (1992). Agricultural statistics by remote sensing in Italy: an ultimate cost analysis. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publication of the E.C. Luxembourg.
- GONZÁLEZ, F., LOPEZ, S., and CUEVAS, J.M. (1991). Comparing two methodologies for crop area estimation in Spain using landsat TM images and ground gathered data. *Remote Sensing of the Environment*, 32, 29-36.
- GOWARD, S.N., MARKHAM, B., DYE, D.G., DULANEY, W., and YANG, J. (1991). Normalized difference vegetation index measurements from the advanced very high resolution radiometer. *Remote Sensing of the Environment*, 35, 257-277.
- HAY, A.M. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9, 1395-1398.
- HOUSTON, A.G., and HALL, F.G. (1984). Use of satellite data in agricultural surveys. *Communications in Statistics Theory and Methods*, 13, 23, 2857-2880.
- VAN LANEN, H.A.J., VAN DIEPEN, C.A., REINDS, G.J., DE KONING, G.H.J., BULENS, J.D., and BREGT, A.K. (1992). Physical land evaluation methods and GIS to explore the crop growth potential and its effects within the European Communities. *Agricultural Systems*, 39, 307-328.
- MEYER-ROUX, J. (1990). Présentation du projet pilote de télédétection appliquée aux statistiques agricoles. Conference on the Application of Remote Sensing to Agricultural Statistics. Office for Publications of the E.C. Luxembourg.
- NEALON, J.P. (1984). Review of the multiple and area frame estimators. U.S. Department of Agriculture, Statistical Reporting Service, Report 80, Washington, D.C.
- SHARMAN, M., and de BOISSEZON, H. (1992). Action IV: de l'image aux statistiques, bilan opérationnel après deux années d'estimations rapides des superficies et des rendements potentiels au niveau Européen. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publications of the E.C. Luxembourg.

Use of Capture-Recapture Techniques to Estimate Population Size and Population Totals when a Complete Frame is Unavailable

K.H. POLLOCK, S.C. TURNER and C.A. BROWN¹

ABSTRACT

We present a formal model based sampling solution to the problem of estimating list frame size based on capture-recapture sampling which has been widely used for animal populations and for adjusting the US census. For two incomplete lists it is easy to estimate total frame size using the Lincoln-Petersen estimator. This estimator is model based with a key assumption being independence of the two lists. Once an estimator of the population (frame) size has been obtained it is possible to obtain an estimator of a population total for some characteristic if a sample of units has that characteristic measured. A discussion of the properties of this estimator will be presented. An example where the establishments are fishing boats taking part in an ocean fishery off the Atlantic Coast of the United States is presented. Estimation of frame size and then population totals using a capture-recapture model is likely to have broad application in establishment surveys due to practicality and cost savings but possible biases due to assumption violations need to be considered.

KEY WORDS: Incomplete frames; Capture-recapture sampling; Angler surveys; Telephone surveys; Access surveys.

1. INTRODUCTION

In classical sampling theory it is assumed that a complete frame exists. There is, at least conceptually, a complete list of population units. It is then possible to draw a probability sample from the population. Estimators of population parameters such as mean or total then have known properties and are easily studied theoretically or numerically. Books on sampling theory such as Cochran (1978) concentrate on this situation and give properties of estimators for common sampling designs such as simple random sampling, stratified random sampling and multi-stage (cluster) sampling.

In practice in surveys of establishments or businesses a complete frame may not exist. Lists of establishments kept by professional associations or government agencies are often incomplete. One approach to tackling this problem is to use the multi-frame approach originally developed by Hartley (1962, 1974). Examples of this approach are the National Agricultural Statistics Service (USDA) farm surveys (Vogel and Kott 1993). These surveys use an incomplete list frame of farms plus an area frame where all farms within a sample unit are enumerated. Therefore the list frame is incomplete while the area frame is conceptually complete. (There is a list of all area units and within each area unit theoretically all farms could be enumerated.)

There are some situations, however, where it may not be possible to use an area frame for practical reasons. All that the researcher may have available may be several

incomplete list frames of establishments. The usual approach in this situation is to merge all the incomplete lists and ignore any remaining incompleteness. Depending on the degree of incompleteness remaining there could be serious negative bias on estimates of population size and population total.

Later we present a formal model based sampling solution to this problem based on capture-recapture sampling. Capture-recapture sampling models are widely used in sampling animal populations (Seber 1982) and also for adjusting the U.S. census for undercoverage (Feinberg 1992). In the simplest case of two incomplete lists we consider "marked" units to be those which occur on both lists and unmarked units to be those which do not occur on both lists. It is easy to estimate total frame size using the Lincoln-Petersen estimator (Seber 1982, p. 59). This estimator is model based with a key assumption being independence of the two lists. Once an estimator of the population size has been obtained it is possible to obtain an estimator of population total for some characteristic if a sample of units has that characteristic measured.

The usual estimator of a population total for simple random sampling without replacement is

$$\hat{Y} = N\bar{y}, \quad (1.1)$$

where N is known and \bar{y} is the mean of the sample, see for example Cochran (1978, p. 21). The variance of \hat{Y} is given by

$$\text{Var}(\hat{Y}) = N^2 \text{Var}(\bar{y}), \quad (1.2)$$

¹ K.H. Pollock, North Carolina State University, Raleigh, NC 27695; S.C. Turner and C.A. Brown, National Marine Fisheries Service, Miami, FL 33149, U.S.A.

where

$$\text{Var}(\bar{y}) = \frac{S^2}{n} \left(\frac{N-n}{N} \right),$$

S^2 is the population variance and $(N - n/N)$ is called the finite population correction factor. The estimator (1.1) is also an unbiased estimator of the population total.

Here our estimator is

$$\hat{Y} = \hat{N}\bar{y}, \quad (1.3)$$

where \hat{N} is obtained from the capture-recapture method.

This means the properties of the estimator (1.3) are more difficult to evaluate because both \hat{N} and \bar{y} are random variables unlike in estimator (1.1) where N is a known quantity. The estimated variance of \hat{Y} here is given by

$$\widehat{\text{Var}}(\hat{Y}) = (\hat{N})^2 \widehat{\text{Var}}(\bar{y}) + (\bar{y})^2 \widehat{\text{Var}}(\hat{N}) + \widehat{\text{Var}}(\bar{y}) \widehat{\text{Var}}(\hat{N}), \quad (1.4)$$

assuming that \bar{y} and \hat{N} are independent and using an exact result due to Goodman (1960). The estimator (1.3) is only an unbiased estimator if \hat{N} and \bar{y} are unbiased estimators of the population size and population mean respectively which is not usually the case in practice. We discuss the estimator (1.3) in the large pelagic fishery survey example in Section 3.

The remainder of the paper is structured as follows. In Section 2 we review the capture-recapture literature to give an overview of the types of models available. In Section 3 we present an example of a sample survey of fishing boats. (We consider a boat analogous to a business establishment). While this example has some unique features we believe it has many features common to other establishment surveys. In the final discussion section we summarize the strengths and weaknesses of using the capture-recapture approach to estimating frame size in establishment surveys. Many of our ideas will require further research.

2. A BRIEF REVIEW OF CAPTURE-RECAPTURE MODELS

It is obviously beyond the scope of this manuscript to review the extensive capture-recapture literature. For more information we recommend Seber (1982), White *et al.* (1982), Pollock *et al.* (1990) and Pollock (1991). Pollock (1991) is a review paper and a good lead into the literature and our treatment in this section follows it very closely. The other references are books and monographs for the serious reader with more time.

Here we briefly discuss the Lincoln-Petersen model for two samples, more general closed population and open

population models for more than two samples, and finally a method which combines closed and open population models in one sampling design. Pollock *et al.* (1990, p. 9) presents a flow chart which shows an overview of the models and how they relate to each other.

2.1 The Lincoln-Petersen Model

This is the oldest, simplest and best known capture-recapture model dating back to Laplace, who used it to estimate the population size of France. It was first used in fisheries by Petersen around the turn of the century. An excellent detailed discussion of this model is given by Seber (1982, Chapter 3).

In the original fisheries setting the method can be described as follows. A sample of M fish is caught, marked, and released. Later a second sample of n fish is captured, of which m are marked. An intuitive derivation of the estimator follows from equating the proportions marked in the sample and the population,

$$m/n = M/N, \quad (2.1)$$

which gives

$$\hat{N} = Mn/m. \quad (2.2)$$

A modified estimator with less bias in small samples is due to Chapman (1951) and is given by

$$\hat{N}_c = [(M+1)(n+1)/(m+1)] - 1. \quad (2.3)$$

An estimate of the variance of \hat{N}_c is given by

$$\widehat{\text{Var}}(\hat{N}_c) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)}. \quad (2.4)$$

See for example Seber (1982, p. 60).

The crucial assumptions of this model are:

- The population is completely closed to additions and deletions,
- all the fish are equally likely to be captured in each sample, and
- marks are not lost or overlooked.

The assumption about closure can be weakened, but even for a completely open population, where this estimator does not apply, a modification of the Lincoln-Petersen estimator is used. The assumption of equal catchability causes problems in most applications. There may just be inherent variability (heterogeneity) in capture probabilities of individual animals due to age, sex or other factors. There may also be a response to initial capture (trap response). In the next section, we consider closed

population models with more than two samples that allow for time variation as well as heterogeneity and trap responses in the animals' capture probabilities. The loss or overlooking of marks can be serious. One way to estimate mark loss is to use two marks (Seber 1982, p. 94).

2.2 Closed Population Models

Closed population models require the assumption that no births, deaths, or migration in or out of the population occur between sampling periods. Therefore, these models are generally used for studies covering relatively short periods of time (e.g., trapping every day for 5 consecutive days). Capture histories for every animal caught are the data needed for obtaining estimates under these models. Important early references are Schnabel (1938) and Darroch (1958), who considered models that assumed equal catchability of animals in each sample.

A set of models that allow capture probabilities to vary due to heterogeneity, (h), trap response (b), time variation (t), (i.e., capture probability for time i differs from that for time j) and all possible two- and three-way combinations of these factors is now available. The eight models [$M(o)$, $M(h)$, $M(b)$, $M(bh)$, $M(t)$, $M(th)$, $M(tb)$, $M(thb)$] were first considered as a set by Pollock (1974) and were more fully developed by Otis *et al.* (1978), White *et al.* (1982), and Pollock and Otto (1983). Otis *et al.* (1978) provided a detailed computer program, CAPTURE, for use with their monograph. An updated version provides estimates for seven of the eight models and a model selection procedure that aids the biologist in choosing a model. The model selection procedure is based on a variety of goodness-of-fits tests. Recently, Menkins and Anderson (1988) have emphasized that the model selection procedure is poor for small populations, unless the capture probabilities are unrealistically high.

2.3 Open Population Models

In many capture-recapture studies, it is not possible to assume the population is closed to additions and permanent deletions. The basic open population model suitable for this situation is the Jolly-Seber model (Jolly 1965; Seber 1965; Seber 1982, p. 196). The Jolly-Seber model allows estimation of population size at each sampling time as well as estimation of survival rates and birth numbers between sampling times. Migration cannot be separated from the birth and death processes without additional information.

The Jolly-Seber model requires the following assumptions:

- (a) Every animal present in the population at a particular sampling time has the same probability of capture,
- (b) every marked animal present in the population immediately after a particular sampling time has the same probability of survival until the next sampling time,

- (c) marks are not lost or overlooked,
- (d) all emigration is permanent, and
- (e) all samples are instantaneous, and each release is made immediately after the sample.

Assumptions (a), (c), and (e) were required under the basic Lincoln-Petersen model described in Section 2.1. Only marked animals are used to estimate survival rates so that, strictly, we do not need to assume equality of marked and unmarked survival rates. In practice however, the biologist will want to use the survival rate estimates to refer to the whole population. The Jolly-Seber model allows for some animals to be lost on capture and hence not returned to the population. The Jolly-Seber model also requires that all emigration is permanent. If animals emigrate and then return to the population this causes so called temporary emigration which is a serious assumption violation and causes major bias in population size estimates.

2.4 Combination of Closed and Open Models

Pollock (1982), Pollock *et al.* (1990) and Kendall (1992) discuss sampling methods which allow the use of closed and open models in one design. One advantage of these methods is that it is possible to allow for unequal catchability whereas in the traditional Jolly-Seber model it is not possible to allow for unequal catchability. They also have the advantage of allowing for temporary emigration of animals.

2.5 Applications of Capture-Recapture Models

Capture-recapture models have obviously been widely applied to wildlife and fishery populations. A variety of novel nonbiological applications of capture-recapture methods have also now appeared. Many authors have applied capture-recapture to estimating the census undercount. (See Feinberg (1992) for a complete bibliography). Cowan, Breakey, and Fischer (1986) used it to estimate the number of homeless people in a city. Greene (1983) has used the method to estimate demographic parameters on criminal populations. Wittes (1974) and Wittes, Colton, and Sidel (1974) have used capture-recapture to estimate numbers of people with illnesses from hospital and other lists. The sampling of elusive human populations using cluster sampling, network sampling, and capture-recapture sampling was discussed by Sudman, Sirken and Cowan (1988).

3. USE OF CAPTURE-RECAPTURE MODELS IN THE LARGE PELAGIC SURVEY

The Large Pelagic survey is an angler survey conducted by the National Marine Fisheries Service using a telephone-access survey design. A sample of fishing boat owners on a list are telephoned to obtain fishing effort (i.e., number

of fishing trips in a period) information. Catch per unit effort (*i.e.*, catch per trip) information is obtained from a second sample of boat owners at access points at completion of their fishing trips. The information from the two surveys is combined to estimate total effort and total catch of important species such as Bluefin Tuna.

A serious problem with this survey is that the list of boat owners used in the telephone survey is very incomplete. Therefore, classical sampling theory which assumes a complete frame of known size (N) is inadequate and has to be modified. The current method of estimating the size of the fishing boat list frame involves combining two lists, (a telephone list with a dockside list) and using the Lincoln-Petersen model. There are questions about whether this is the best approach. For example, it might be possible to combine more than two lists and if so then we could use the closed or open population models reviewed in Sections 2.2 and 2.3. However, we defer those questions and begin by reviewing and evaluating the current method as an example to illustrate the potential usefulness of the approach to other establishment surveys.

3.1 The Lincoln-Petersen Model

3.1.1 Estimation of Frame Size (N)

Under the current method the “marked” boats (M) are those on the master list which is primarily derived from previous telephone interviews. The recapture sample is carried out dockside at gas pumps and the total number of boats intercepted (n) is checked to see which ones are “marked” (m) (*i.e.*, on the original master list). Equation 2.3 can then be used to provide an estimator of the frame size (N). Let us now consider the assumptions of this model and what effect violations might have on the bias of the estimator of N .

Closure

This assumption is likely to be violated. Fishing boats may be on the master list and then no longer take part in the fishery (losses). New fishing boats may join the fishery while it is in progress (gains). Ideally a separate estimate of frame size should be obtained for each two week time period. The advantage of using the Lincoln-Petersen closed model estimator is its simplicity and practicality. Biases in the estimator due to lack of closure could be either positive or negative.

Currently it is not known how the fishing fleet size is likely to change during the fishing season. A multiple capture-recapture sampling design would allow use of the Jolly-Seber model to estimate the fleet size during each period. Examination of these estimators and the survival rate and recruitment number estimators will enable us to evaluate the validity of the closure assumption. At the moment we can only make conjectures.

Equal Catchability

Violation of the assumption of equal catchability may be due to either inherent heterogeneity of capture probabilities between individuals or “trap response” where individuals that are marked have higher or lower capture probabilities than unmarked individuals. In either situation when the individuals on the lists are fishing boats we believe there is a potential for heterogeneity of capture probabilities among fishing boats. If heterogeneity is operating across both samples, individuals “caught” on the first list will tend to be those with high capture probabilities and therefore they will more likely to be “caught” again on the second list. This means that the proportion marked in the second sample (list) will be too high and the estimator of N will be negatively biased. Note that this intuitive argument makes clear it is not heterogeneity per se which is the problem but the positive correlation of capture probabilities between the two samples. Another way of stating the equal catchability assumption is that capture probabilities in the two samples are independent. One method of attempting to achieve independence of the capture probabilities in the two samples is to use totally different sampling schemes for the two samples. This is why we recommended earlier that one sample list be based on the telephone interviews and the other on dockside interviews. However, we do suspect that there is still another heterogeneity and lack of independence in capture probabilities. We believe that fishing boats which take a very active part in the fishery are more likely to be on any lists gathered (telephone or dockside). This heterogeneity will cause a negative bias on the estimate of frame size but we have no idea of the degree of this negative bias. A more complete discussion of heterogeneity and independence of samples is given by Seber (1982, p. 86).

Marks Lost or Overlooked

The situation here is a little confusing. At first one might think that in this application there is not a way that a mark could be lost or overlooked. However, this assumes that all boats have distinct names or that if boats do have the same name there is additional information like captain's name which makes all individuals on the lists unique. If there is any problem with lack of uniqueness it may not be clear whether a marked boat has been recaptured or not. Another related point is that agents may make errors in the records which make it hard to match up a recapture with the original record. A standard operating procedure is being developed and documented to minimize these kinds of errors in the future.

3.1.2 Estimation of Total Effort and Total Catch

Total Effort (E) (*i.e.*, the total number of fishing trips taken in a defined period) is estimated by

$$\hat{E} = \hat{N}\bar{e}, \quad (3.1)$$

where \hat{N} is the frame size (Fleet Size) estimate and \bar{e} is the mean fishing effort (*i.e.*, average number of fishing trips taken) obtained from the telephone sample. The evaluation of the properties of this estimator is more difficult than when N is known because both \hat{N} and \bar{e} are random variables. We suspect that \bar{e} is biased high because fishing boats that do not fish much are less likely to be on the list. Unfortunately we cannot say that \hat{N} will always be biased high or low. All three of the assumption violations discussed in 3.1.1 could be important (closure, heterogeneity, and mark loss) and it is not clear what direction the overall bias on \hat{N} would take. The only possible approach is to use simulation with a variety of different scenarios for assumption violations. Using equation (1.4) the estimated variance of \hat{E} is given by

$$\widehat{\text{Var}}(\hat{E}) = (\hat{N})^2 \widehat{\text{Var}}(\bar{e}) + (\bar{e})^2 \widehat{\text{Var}}(\hat{N}) + \widehat{\text{Var}}(\bar{e}) \widehat{\text{Var}}(\hat{N}). \quad (3.2)$$

Total catch (C) is estimated by $\hat{C} = \hat{E}\bar{c}$ where \hat{E} is the estimated total fishing effort and \bar{c} is the average catch per unit effort calculated from the dockside interviews. Properties of this equation are likely to be subject to similar concerns as equation (3.1) and again simulation could be very useful.

3.1.3 Illustration of the Method

In this section we present the frame size estimates and total effort estimates for the Virginia Bluefin tuna fishery in part of 1992. These estimates are a part of a larger survey which covered the east coast of the U.S. from North Carolina to Massachusetts. The estimates are separate for charter boats and private boats.

Frame Size Estimates

Lists of unique private boats and charter boats were compiled mainly by telephone interviews from previous seasons. During the current 1992 season “marked” and “unmarked” boats were captured at gas pumps before or after fishing trips.

For private boats the list size was $M = 335$ boats before the season. A sample of $n = 374$ boats were contacted at gas pumps and of those $m = 49$ were marked. The Chapman estimator is $\hat{N}_c = 2,519$, $\widehat{SE}(\hat{N}_c) = 303.08$ and relative $\widehat{SE} = 0.12$.

For charter boats the list size was $M = 47$ before the season. A sample of $n = 31$ boats were contacted at gas pumps and of those $m = 13$ were marked. The Chapman estimator is $\hat{N}_c = 109$ with $\widehat{SE}(\hat{N}_c) = 17.88$ and relative $\widehat{SE} = 0.16$.

Total Effort Estimates

Total effort and total catch were estimated in weekly waves. Here we just illustrate the calculations for the week of the 8th to the 14th of June 1992 for total effort.

Total Effort – Private Boats

$\hat{N}_c = 2,519$ boats, $\widehat{\text{Var}}(\hat{N}_c) = 91,856.4706$, $\bar{e} = 0.15108$ trips per interview, $\widehat{\text{Var}}(\bar{e}) = 0.001242$ and $\widehat{SE}(\bar{e}) = 0.0352$. Using these estimates we obtain

$$\hat{E} = \hat{N}_c \times \bar{e} = 2,519 \times 0.15108 = 380.57 \text{ trips,}$$

$$\widehat{\text{Var}}(\hat{E}) = \widehat{\text{Var}}(\bar{e})(\hat{N}_c)^2 + \widehat{\text{Var}}(\hat{N}_c)(\bar{e})^2 +$$

$$\widehat{\text{Var}}(\hat{N}_c) \widehat{\text{Var}}(\bar{e}) = 10,091.6633, \text{ and}$$

$$\widehat{SE}(\hat{E}) = 100.45.$$

It is useful to also calculate the variance of total effort assuming that the frame size were known. In this case it is $\widehat{\text{Var}}(\hat{E}) = 7,780.9384$ with $\widehat{SE}(\hat{E}) = 88.77$ and this shows that 89% of the standard error of the Total Effort estimate is due to variation in average effort and only 11% is due to estimation of frame size.

Total Effort – Charter Boats

For charter boats $\hat{E} = 59.95$ trips with $\widehat{\text{Var}}(\hat{E}) = 512.5100$ and $\widehat{SE}(\hat{E}) = 22.64$.

The variance of the Total Effort estimate assuming the frame size is known is $\widehat{\text{Var}}(\hat{E}) = 404.8926$ with $\widehat{SE}(\hat{E}) = 20.12$. Again 89% of the standard error of the Total Effort estimate is due to variation in average effort and only 11% is due to estimation of frame size.

3.2 More Than Two Lists

In Section 2 we indicated that there are a lot more modeling possibilities if one has multiple (greater than 2) lists. Here we consider closed and open population models for the more general case. We foresee the sampling scheme as follows. Before the start of the fishing season there would be a preliminary sample to establish a list (either telephone or dockside). During each time period (say two weeks) there would be an additional list compiled using a telephone or dockside survey. Now each individual boat would have a capture history which would indicate which lists it appeared on. (Suppose we have five time periods then a capture history of 1 1 1 0 1 would indicate a boat appeared on the lists in all except the fourth time period).

The structure of the sample and the population would therefore be as in Table 1. The first question that has to be addressed is whether we need to use closed or open population models. The obvious way to proceed is to fit the Jolly-Seber open population model first and use it to evaluate the closure assumption.

Table 1

Structure of the Population Under an Open Population Model*

Period	Pre-season List	Season Lists (<i>e.g.</i> , every two weeks)					
		0	1	2	3	.	<i>k</i>
Marked Population Sizes	M_0	M_1	M_2	M_3	.	.	M_k
Total Population Sizes	N_0	N_1	N_2	N_3	.	.	N_k

* Marked and Total Population Sizes are shown for the whole study.

3.2.1 Open Population Models

Under the Jolly-Seber model previously discussed in Section 2.3 the following parameters are identifiable (Table 2). Notice that it is possible to estimate the number of fishing boats in the fleet at each time in the season except the last (*i.e.*, \hat{N}_k cannot be estimated). One advantage of applying the model in this fashion with a preseason list is that any concerns with the preseason list due to it being out of date are taken care of by the model allowing for additions and deletions before the season begins. One disadvantage of the Jolly-Seber Model is increased complexity. Now each time period has its own frame size and there are also survival and recruitment parameters to estimate. Sometimes these parameter estimates have poor precision unless sample sizes are large. Another disadvantage of the Jolly-Seber model is that it does require the assumption of equal catchability.

Table 2

Structure of the Jolly-Seber Open Population Model*

Period	Preseason	Season						
	0	1	2	3	.	.	$k-1$	k
Marked Population	$(M_0 \equiv 0)$	\hat{M}_1	\hat{M}_2	\hat{M}_3	.	.	\hat{M}_{k-1}	-
Total Population	-	\hat{N}_1	\hat{N}_2	\hat{N}_3	.	.	\hat{N}_{k-1}	-
Survival Rate	$\hat{\rho}_0$	$\hat{\rho}_1$	$\hat{\rho}_2$.	.	$\hat{\rho}_{k-2}$.	-
Recruitment No.		\hat{B}_1	\hat{B}_2	.	.	\hat{B}_{k-2}	.	-

* Identifiable parameter estimators are shown for Marked Population Sizes, Total Population Size, Survival Rate and Recruitment Number.

Another important question about the use of the Jolly-Seber model is what is called “temporary emigration.” A fishing boat might leave the fishery for some periods and then return. The Jolly-Seber model makes the assumption that fishing boats which leave do not return. This issue needs further investigation. Use of the robust design (*i.e.*, combination closed and open models) allows for temporary emigration. This would necessitate having two lists obtained close together in each period.

3.2.2 Closed Population Models

If the Jolly-Seber model estimates of “survival” and “recruitment” suggest population closure (*i.e.*, N constant) then the general closed population models reviewed in Section 2.2 could be applied. The advantages are increased precision of \hat{N} due to the use of more lists and increased robustness of \hat{N} to unequal catchability. The disadvantage is primarily an increase in complexity.

4. DISCUSSION

4.1 Methods of Dealing with Incomplete List Frames

(i) Complete the List Frame

The advantage is that the survey researcher has a complete frame and does not have to generalize results for an estimated frame size. The disadvantage is the cost and possible impracticality of completing the list frame.

(ii) Use an Area Frame

The advantage is that one only has to enumerate the establishments in the areas to be sampled. The disadvantage is possible inefficiency if businesses are sparse in each large area.

(iii) Using List and Area Frame (Multi-Frame Approach)

The advantages are obviously increased precision and having all establishments covered. The disadvantage could be expense and impracticality.

(iv) Use of Capture-Recapture to Estimate List Frame Size

The advantage is having a practical method of lower expense than the first three approaches listed above. The disadvantages are potential bias if the assumptions of the capture-recapture method are violated and having to include variation due to frame size estimation in variance estimates of population total estimates.

4.2 Capture-Recapture Estimation of Frame Size

In this section we consider model assumptions, precision of estimates, estimation of population totals and the special problems in more complex sampling designs when the capture-recapture approach to frame size estimation is used.

Model Assumptions

(i) Closure

Can the frame size be considered constant so that the closed population models be used? This will depend on whether the survey is just a snapshot at a single time point or whether a series of surveys over time are required. It will also depend on how quickly establishments go out of business and how quickly new ones arise. We suspect there will be the need for use of closed and open population models depending on the establishments being studied.

There is also the question of temporary emigration where establishments go out of the frame and then come back in again. This was considered a potential problem in the fishing boat example because boats could go inactive and then become active again. This may also be a problem in some other establishment surveys if establishments go in and out of business frequently and keep the same name when they come back into business.

(ii) "Unequal Catchability" and Independence of Lists

As we discussed earlier ideally the lists used should be independent so that the estimates of frame size are unbiased. In practice it may not be easy to find two or more independent lists.

(iii) Mark Loss-Unique Identification of Establishment

Establishment names need to be unique and unmis-takable or matches on different lists may be missed or mistaken. This was a problem in the fishing boat example in earlier years. We suspect this will not be such a big problem in most establishment surveys.

Precision of Estimates

The lists used need to be of sufficient size that the precision of the frame size estimate (\hat{N}) is adequate. Seber (1982, p. 96) discusses the Lincoln-Petersen estimate in detail and presents graphics of sample sizes required for various levels of precision. Pollock *et al.* (1990) presents sample size information for the open population models.

Estimation of Population Totals

Once the estimate of frame size is obtained then that estimate will often be combined with a sample mean to obtain an estimate of a population total ($\hat{Y} = \hat{N}\bar{y}$). The estimate of population total is subject to possible bias and additional variance because \hat{N} is estimated. The estimate may also be biased because \bar{y} is not based on a random sample of the complete frame.

More Complex Sampling Designs

In this paper we have emphasized estimation of frame size in simple random sampling using the capture-recapture method. Further questions arise if more complex sampling designs are used. For example in stratified designs the question would arise of whether to estimate frame size in each stratum separately or to estimate the total frame size and then apportion it to the strata assuming equal probabilities of different strata on the incomplete lists. There is also the more complex question of how to estimate frame size in multi-stage sampling designs. This is obviously an area that needs future research.

ACKNOWLEDGEMENTS

The authors wish to thank the editor, associate editor and two anonymous referees for helpful comments which have greatly improved the paper.

REFERENCES

- CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with application to zoological census. *University of California Publication in Statistics*, 1, 131-160.
- COCHRAN, W.G. (1978). *Sampling Techniques* (Third Edition). New York: John Wiley and Sons.
- COWAN, C.D., BREakey, W.R., and FISCHER, P.J. (1986). The methodology of counting the homeless. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-175.
- DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.
- FIENBERG, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18, 143-154.
- GOODMAN, L.A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-713.
- GREENE, M.A. (1983). Estimating the size of the criminal population using an open population approach. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-13.
- HARTLEY, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, C, 36, 99-118.
- JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52, 225-247.
- KENDALL, W.L. (1992). Robust Design in Capture-Recapture Sampling: Modelling Approaches and Estimation Methods. Unpublished Ph.D. dissertation, North Carolina State University, Biomathematics Program.
- MENKINS, G.E., and ANDERSON, S.H. (1988). Estimation of small mammal population size. *Ecology*, 69, 1952-1959.
- OTIS, D.L., BURNHAM, K.P., WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-125.
- POLLOCK, K.H. (1974). The Assumption of Equal Catchability of Animals in Tag-Recapture Experiments. Unpublished Ph.D. dissertation, Cornell University, Biometrics Unit.
- POLLOCK, K.H. (1982). A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46, 752-757.
- POLLOCK, K.H., NICHOLS, J.D., HINES, J.E., and BROWNIE, C. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107, 1-97.
- POLLOCK, K.H., and OTTO, M.C. (1983). Robust estimation of population size in closed animal populations from capture-recapture experiments. *Biometrics*, 39, 1035-1049.
- POLLOCK, K.H. (1991). Modelling, capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. *Journal of the American Statistical Association*, 86, 225-238.
- SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.

- SEBER, G.A.F. (1965). A note on the multiple recapture census. *Biometrika*, 52, 249-259.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters* (Second Edition). New York: MacMillan.
- SUDMAN, S., SIRKEN, M.G., and COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-995.
- VOGEL, F.A., and KOTT, P. (1993). Multiple frame establishment surveys. *Proceedings, International Conference on Establishment Surveys*.
- WHITE, G.C., ANDERSON, D.R., BURNHAM, K.P., and OTIS, D.L. (1982). *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos, NM: Los Alamos Laboratory.
- WITTES, J.T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 79-93.
- WITTES, J.T., COLTON, T., and SIDEL, V.W. (1974). Capture-recapture method for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.

Questionnaire Design for Business Surveys

A.R. GOWER¹

ABSTRACT

This paper provides an overview of important considerations that should be taken into account when developing and designing questionnaires for business surveys. These considerations include the determination of objectives and data requirements, consultation with data users and respondents, and methods for testing questionnaires. In developing and designing business survey questionnaires, focus groups and cognitive research methods help the researcher to identify potential sources of measurement error and to understand the response process that respondents go through in completing the questionnaires. Examples of focus groups and cognitive research undertaken by Statistics Canada are provided.

KEY WORDS: Questionnaire testing; Focus groups; Cognitive research.

1. INTRODUCTION

There are many types of business survey questionnaires. Typically, a business survey questionnaire collects information about a company's employees, its inventories, inputs, products, sales, and finances. It may also involve the collection of information related to market research or client satisfaction.

Business surveys are conducted by mail or administered by an interviewer in person or over the telephone. Follow-ups to mail surveys are often conducted by telephone. New data collection technologies for business surveys involve computer-assisted interviewing, fax machines, touchtone self-response, and the electronic transmission of data.

As in other types of surveys, questionnaires play a central role in the data collection process in a business survey. They have a major impact on data quality and on the image that a survey organization projects to its respondents.

The purpose of this paper is to provide an overview of questionnaire design for business surveys. The paper discusses important considerations such as the determination of objectives and data requirements, consultation with data users and respondents, the nature and concerns of business survey respondents, and methods for testing questionnaires.

In developing and designing business survey questionnaires, it is especially important to understand the response process that respondents go through in completing the questionnaires. Therefore, this paper emphasizes the effectiveness of using focus groups and cognitive research techniques to develop and test business survey questionnaires. Examples of focus groups and cognitive research that have been carried out by the Questionnaire Design Resource Centre of Statistics Canada are provided.

2. BUSINESS SURVEY QUESTIONNAIRES

A well-designed questionnaire in a business survey should collect data efficiently, with a minimum number of errors. Moreover, questionnaires should facilitate the coding and capture of data. They should minimize the amount of editing and imputation that is required. They should also lead to an overall reduction in the cost and time associated with data collection and processing (Statistics Canada 1994).

There are many considerations that apply to the development and design of business survey questionnaires. One key consideration is the nature of the respondent population. Business survey respondents answer in their role as employers or employees of a business. How a questionnaire is completed depends on the position and level of responsibility that the respondent holds in the business organization or company. Therefore, it is critical to identify the most appropriate person to provide the information in a business survey.

Response burden is a very real concern for business survey respondents. It depends on the number of questions that are asked, the time required to complete the questionnaire, and the effort that respondents put into searching or manipulating other data sources to provide the information in the format requested.

Businesses vary in size. Large businesses may have employees whose responsibilities include completing government and survey forms. In small businesses, respondents are often the owners or office managers who may not have as much time or flexibility in their schedules to complete the questionnaire.

Information provided by respondents in business surveys typically involves the use of records or other information systems. Questionnaires often contain technical or professional terminology associated with providing financial or administrative data.

¹ A.R. Gower, Questionnaire Design Resource Centre, Statistics Canada, Ottawa, Ontario, K1A 0T6.

Another consideration is the confidentiality and sensitivity of the information that the questionnaire is collecting. In many cases, businesses are concerned about providing confidential financial information that they do not want to reveal to competitors, governments or any other party. Therefore, assurances of confidentiality should be provided. All necessary arrangements should be made for the proper handling and custody of data in order that the confidentiality of information is ensured.

3. THE RESPONSE PROCESS IN BUSINESS SURVEYS

The model of the response process is well-known for household surveys. Answering these types of questions involves comprehension, retrieval, thinking/judging, and responding (Tourangeau 1984). Respondents must first understand the question. They then search their memories to retrieve the requested information. After retrieving the information, they think about what the correct answer to the question might be and how much of that answer they are willing to reveal. Only then do they give an answer to the question.

A corresponding response model for business surveys has also been developed (Edwards and Cantor 1991). Although the business survey model is similar to the household survey model, there are differences. The major difference is that business survey respondents must normally access one or more external sources of information such as financial or administrative records.

The ability of respondents to retrieve the requested information depends upon their familiarity with and understanding of the external source of information. They must also understand the relationship between the survey questions and the external data source. Multiple sources of information may add to the difficulty or complexity of this task. Further complexities may be introduced if the respondent has to consult another individual who can provide the requested information and who, in turn, may have to use one or more data sources (Gower and Nargundkar 1991).

4. DEVELOPMENT AND TESTING OF BUSINESS SURVEY QUESTIONNAIRES

There are several basic steps that are involved in developing and testing business survey questionnaires. These steps are discussed below.

4.1 Determination of the Objectives and Data Requirements

A document should be prepared that provides a clear and comprehensive statement of the survey objectives,

data requirements, and the data analysis plan. This document is a necessary step that leads to the determination of the variables to be measured, the survey questions, and the response alternatives.

When designing the questionnaire, it is important to determine and understand the rationale for each question, how the information will be used, and whether the questions will be good measures of what is required.

4.2 Consultation with Clients, Data Users, Subject Matter Experts, and Respondents

In formulating objectives and data requirements, consultation should take place with clients and data users to fully understand their requirements and expectations. Subject matter experts should be contacted for advice and guidance.

If possible, the survey researcher should consult members of the survey population. This will help identify issues and concerns that are important to respondents, and may affect decisions regarding the content of the questionnaire. In addition, consultation with respondents will identify the language and terminology that respondents themselves use and will help clarify terminology, concepts and definitions.

4.3 Previous Questionnaires

Examining questions that were used in other surveys on the same or a similar topic provides a useful starting point in formulating the questions and response categories. In some situations (*e.g.*, for comparing data over time), the same questions may be used. The researcher should ensure that the questions are phrased so as to provide valid, consistent, and effective measures of the variables of interest.

4.4 The Use of Focus Groups in Developing Questionnaires

A *focus group* is an informal discussion of a selected topic involving participants who are chosen from the survey population. It provides insights into the attitudes, opinions, concerns, and experiences of the participants. A focus group is led by a moderator who is knowledgeable about group interviewing techniques and the purpose of the discussion.

Focus groups provide the opportunity to consult respondents, data users, and interviewers. In the early stages of developing a questionnaire, focus groups are used to develop the survey objectives and data requirements, to identify salient research issues, and to clarify definitions and concepts.

Focus groups are also useful in testing and evaluating questionnaires (see 4.6 below). They are used to evaluate respondents' understanding of the language and wording used in questions and instructions, and to evaluate alternative question wordings and formats.

Recruiting participants from businesses poses unique challenges for focus groups. Monetary incentives or honoraria that are usually offered to focus group participants (currently in the order of \$30 to \$50 each) may not be appropriate for business people. Assurances of confidentiality and emphasis on the importance of the survey and their participation in the study are more meaningful. Another type of incentive that may be offered is a donation to a non-profit organization of the participant's choice. Statistics Canada often gives focus group participants a copy of a publication that is of interest to them.

Focus groups vary in size from 6 to 12 persons. The optimum size is 7 or 8 persons for business participants, although smaller groups with 4 or 5 people (called mini focus groups or mini groups) are sometimes held. Because of difficulties in finding participants from businesses, focus groups should be conducted at a time that is convenient to the participants. For business people, focus groups are often held during working hours. Focus groups are audio-recorded, and are viewed by observers in an adjoining room behind a one-way mirror. Participants are fully informed that audio-recording is taking place and that they are being observed.

4.5 Considerations in Drafting the Questions

Many considerations go into writing the questions and developing the response categories. It is important to keep in mind the objectives and data requirements as well as how the information will be collected and processed. The questions must relate to the information needs. They must be addressed to the right people in the organization or company.

The method of data collection will determine how the questions and response categories will be formulated. The question wording must be clear, and they must be ordered in a logical sequence. The questions must be designed to be easily understood and accurately answered by respondents. Response categories and time reference periods should be compatible with the business's record-keeping practices; however, this is often difficult to achieve.

The layout of the questionnaire should be attractive. The questionnaire should be *respondent-friendly* and, if administered by an interviewer over the telephone or in person, it should be *interviewer-friendly*.

The questionnaire should appear professional and "business-like". When designing the questionnaire, it should be kept in mind that businesses are asked to complete many forms and questionnaires. Completing them is not a priority. Research conducted by Statistics Canada's Questionnaire Design Resource Centre has shown that typical reactions from businesses to questionnaires are:

- "I complete the shortest form first."
- "Is completion mandatory?"
- "Is there a return deadline?"

In one Statistics Canada study (Gower and Zylstra 1990), a respondent commented that if the answer to these last two questions is "no," then "I put [the questionnaire] in my *maybe I'll get to it someday* basket!"

Respondents frequently question the value of information to themselves and to other users. Some like to receive feedback about the survey. Therefore:

- Explain why it is important to complete the questionnaire.
- Ensure that the value of providing information is made clear to respondents.
- Explain how the survey data will be used.
- Explain how respondents can access the data.

The instructions that go with the questionnaire also require attention. Research carried out by the Questionnaire Design Resource Centre has repeatedly shown that respondents read only what they *think* is necessary to read. They read the boldface print first, and then decide whether they should read further. Respondents rarely read the instructions, and usually proceed directly to the questions. They refer to the instructions only when they *think* they need help. As a result, respondents may miss important instructions and definitions. Errors in reporting are often due to a lack of clear instructions and due to respondents not reading them or not understanding them (*e.g.*, what to include or exclude). Therefore:

- Ensure that instructions are short and clear.
- Tell the respondent where to find the instructions.
- Provide definitions at the beginning of the questionnaire or in specific questions as required.
- Use **boldface print** or underlining to emphasize important items such as the reference or reporting period.
- Specify "include" or "exclude" in the questions and items themselves (not in separate instructions).

Other considerations that should be taken into account in designing business survey questionnaires include:

- Consistency of terminology, questions and response categories with standard concepts and definitions.
- Nature of the respondent population such as record-keeping practices and language ability.
- Availability of the data.
- Response burden.
- Complexity of the data to be collected.
- Comparability of results with other surveys.
- Data reliability.
- Nonresponse.

The design of the questionnaire should also take into account any administrative requirements of the survey organization. For example, Statistics Canada's policy on informing survey respondents (Statistics Canada 1986) requires that key information be explained to respondents. They must be informed about the main purpose(s) of the

survey, the major intended uses of the data, the requirement to respond (compulsory or voluntary), confidentiality protection, and any joint collection or data sharing agreements. At Statistics Canada there are also other administrative or legal requirements. For example, the Official Languages Act of Canada requires that questionnaires be made available to respondents in both official languages (*i.e.*, English and French).

4.6 The Use of Cognitive Methods in Testing Questionnaires

Questionnaire testing is essential to developing effective questionnaires that collect useful and accurate data. Cognitive research methods, sometimes referred to as qualitative testing, are especially useful in testing questionnaires.

Cognitive methods provide the means to examine respondents' thought processes as they answer the survey questions. They are used to ascertain whether or not respondents understand what questions mean and thus help assess the validity of questions and identify potential sources of measurement error. Cognitive methods also provide the opportunity to evaluate the questionnaire from the respondent's point of view. They focus on issues such as comprehension and reactions to the form. This brings the respondent's perspective directly into the questionnaire design process. The use of cognitive methods leads to the design of respondent-friendly questionnaires that can be completed easily and accurately.

In business surveys, cognitive methods are used to investigate the relationship between the respondent and the external information source. They are also used to study the influence that this data source has on the response process. These methods provide the means to assess the compatibility of question wording, time reference periods, and response categories with the business's record-keeping practices.

Cognitive testing methods (Gower 1993) include:

- *In-depth interviews*: The technique involves one-on-one interviews (sometimes called retrospective think-aloud interviews). For a mail questionnaire, respondents first complete the questionnaire as they normally would. An interviewer observes the process, noting the sequence in which the questions are answered, reference made to instructions, and the types of records or other persons consulted. The interviewer also notes the time required to complete sections, and corrections or changes made to responses.

The interviewer then conducts the in-depth interview and obtains information about the respondent's experiences and impressions in completing the form. The follow-up discussion typically involves a question-by-question review of the questionnaire with the respondent to discuss any problems or difficulties that were encountered

while completing the form. The interviewer probes to see how the terms and concepts were interpreted by the respondents, how and why they chose the responses, and how information was recalled.

For an interviewer-administered questionnaire, the questions are first asked by an interviewer either in person or by telephone. The in-depth follow-up discussion takes place following this first interview.

- *Concurrent think-aloud interviews*: These are also conducted one-on-one. The respondent is asked to "think aloud" while answering the questions, commenting on each question and explaining how the final response was chosen. The observer may probe the responses to get more information about a particular statement or to clarify the process through which a response was chosen.

The success of the concurrent think-aloud interview technique depends on the respondent's ability and willingness to articulate and express thoughts aloud. The observer may sometimes have to help the respondent in this task by gentle prompts such as: "what question are you answering now?", "what are you thinking now?", "please explain how you chose the answer", or other probes to clarify the respondent's thoughts. When a respondent is reluctant to verbalize thoughts, the observer may decide that the better approach is to handle the interview as an in-depth interview and proceed accordingly.

Think-aloud interviews are very useful in obtaining respondents' reactions to questionnaires. They are especially helpful in identifying areas of the questionnaire where respondents have difficulty. They also help the researcher understand the process through which the questionnaire is completed.

- *Focus groups*: As described in 4.4, focus groups are used to evaluate respondents' understanding of the language and wording used in questions and instructions. The questionnaire is usually administered before the focus group session, in person, over the telephone or on a self-completion basis.

During the focus group session, the moderator reviews the questionnaire with the participants and discusses any problems or difficulties that they may have encountered when completing the form. Focus groups stimulate and encourage thoughtful analysis of the questionnaire during group discussions of individual participants' comments. They are especially useful in providing suggestions and recommendations for improvements.

- *Paraphrasing*: Paraphrasing is used in one-on-one interviews and focus groups. Respondents are asked to repeat the question in their own words, or to explain the meaning of terms and concepts that are used in the survey questions and instructions.

Paraphrasing helps determine whether respondents read and understand the instructions and questions correctly. Paraphrasing is especially helpful in identifying question wording that is too complex or confusing. It also identifies situations where respondents do not comprehend all the important components of the question (*e.g.*, the reference period).

4.7 Pretesting

Pretesting is a fundamental step in developing a questionnaire. It usually involves a small number of field interviews that are carried out to identify problems with a questionnaire. The entire questionnaire or only a portion of it may be tested.

Pretests are useful for discovering poor question wording or ordering, errors in questionnaire layout or instructions, and problems caused by the respondent's inability or unwillingness to answer the questions. Pretests are also used to suggest additional response categories that can be pre-coded on the questionnaire. Pretests provide a preliminary indication of the interview length and refusal problems.

The pretest sample can range in size from 20 to 100 or more respondents. If the main purpose of the pretest is to discover wording or sequencing problems, only a small number of interviews may be required. More interviews (50 to 100) are necessary to determine pre-coded answer categories for open-ended responses. Respondents for pretests are usually selected purposively rather than randomly.

The questionnaire for a pretest should be administered in the same way as planned for the main survey (*e.g.*, interviewer-administered in person or by telephone). A pretest of a mail questionnaire is more effective if interviewers are used. Interviewers can be used to deliver the questionnaire and, later, to discuss any problems. The questionnaire designers should observe as many pretest interviews as possible.

Pretesting is not as effective as cognitive methods in evaluating respondents' understanding and the difficulty of the response task. Pretesting only indicates whether there is a problem. Without further investigation, it does not identify why there is a problem nor how it can be corrected.

Debriefing sessions with interviewers often occur in conjunction with a pretest. Interviewers involved in a pretest can identify important problem areas where the questionnaire can be improved. When existing questionnaires are redesigned, it is useful to consult interviewers to get their input into the redesign process. Interviewers have excellent insights into the logistics of administering the questionnaire and how it affects respondent cooperation.

Behavioral coding also can be conducted at the time of pretesting. The interview is audio-recorded, following which the interviewer and respondent behaviours during

the interviewer-respondent interaction are coded and analyzed. Behavioral coding provides a systematic and objective means of examining the effectiveness of the questionnaire. It also helps to identify problem areas such as an interviewer failing to read the question as worded or a respondent asking for clarification of the question or response task.

4.8 Formal Testing Methods

Formal testing methods are quantitative in nature. They are designed to provide a statistical evaluation of how the questionnaire performs. Pilot studies and split sample testing are two commonly used types of formal testing methods. These methods are more suitable for large scale and continuing surveys because of the significant cost involved in implementing them and analyzing the results.

A *pilot study* is conducted to observe how all the survey operations, including the administration of the questionnaire, work together in practice. A pilot study is a "dress rehearsal". It duplicates the final survey design on a small scale from beginning to end, including data processing and analysis. It allows the survey researcher to see how well the questionnaire performs in relation to all other parts of the survey. There are some problems that can only be identified when all phases of the survey are tested together. For example, typographical errors and problems with question wording or concepts that need further clarification may be identified during interviewer training. The data processing phase may reveal keying problems with the precoded item numbers and/or answer categories (DeMaio 1983).

Normally, the questionnaire should be thoroughly pre-tested before a pilot study takes place. A pilot study is usually not the time to try out new questions or approaches. If previous testing has been carried out, it is unlikely that the pilot study will result in major changes to the questionnaire. The pilot study, however, does provide the opportunity to fine-tune the questionnaire before its use in the main survey (DeMaio 1983).

Split sample testing is conducted to determine the "best" of two or more alternative versions of the questionnaire. Split sample testing is also referred to as a "split ballot" or "split panel" experiment. It involves an experimental design that is incorporated into the data collection process. A split sample test can be designed to investigate issues such as question wording, question sequencing, the location of sensitive items, and data collection procedures. In a simple split sample design, half of the sample is selected at random and might receive one experimental treatment and half, the other. In a test that involves two experimental treatments, a 2×2 factorial design might be used with each of the two treatments in each experiment being tested on half of the sample (DeMaio 1983).

A split sample design can also be used in continuing surveys that assess trends over time and compare results across surveys. In these types of surveys, there often is a concern that any change in the questionnaire or procedures may affect other data items besides the items being added or revised. In these cases, a split sample design may be used with a random sample of the respondents receiving the “old” questionnaire and the rest, the “new” questionnaire. Comparisons with earlier data can still be made by using the old questionnaire for most or part of the sample (DeMaio 1983).

4.9 Review and Revision of the Questionnaire

The questionnaire should be reviewed by someone outside the project team. Reviewers could include subject matter experts or persons who have experience in designing questionnaires. A review can take place at any or all stages of the questionnaire development process, causing revisions in the questions and response categories.

Questionnaire design is an iterative process. Throughout the whole process of questionnaire development, revision and testing, changes will be made continually to improve the questionnaire. Objectives and information requirements are stated, evaluated and decided upon, data users and respondents are consulted, proposed questions are drafted and tested, questions are reviewed and revised, until a final questionnaire is developed.

5. APPLICATION OF FOCUS GROUPS AND COGNITIVE RESEARCH METHODS TO TEST BUSINESS SURVEY QUESTIONNAIRES

Statistics Canada has found that focus groups and cognitive research methods are very useful in developing and testing business survey questionnaires. These methods provide the opportunity to understand the cognitive processes involved in formulating responses to survey questions. They bring the respondent’s perspective directly into the questionnaire design process and lead to the design of respondent-friendly questionnaires (Gower and Nargundkar 1991).

Statistics Canada’s applications of focus groups and cognitive research methods for business surveys include the developing and testing of questionnaires for the following surveys:

- Survey of Employment, Payrolls and Hours (Bureau 1991; Goss, Gilroy and Associates Ltd. 1989; Goss, Gilroy and Associates Ltd. 1990).
- Census of the Construction Industry (Gower and Zylstra 1990; Price Waterhouse Management Consultants 1990).
- Wholesale and Retail Trades Survey (Noonan 1992).

- National Training Survey (Kennedy and de Groh Consultants 1992; D.R. Harley Consultants Limited 1993).

These studies involved the application of one or more of the following methods: focus groups, in-depth interviews, concurrent think-aloud interviews, and paraphrasing. All studies were carried out under the coordination and general direction of Statistics Canada’s Questionnaire Design Resource Centre (Gower 1991).

Each of the studies has demonstrated the importance of and benefits to be gained from consulting with members of the target population before developing and finalizing the questionnaire. The studies have provided valuable insights into the response process and have identified various factors that contribute to measurement errors in business surveys. These factors include the respondents’ perceived value of the information, their perception of response burden, the compatibility of questions with their record-keeping practices, the placement and use of instructions, the availability of data, and the complexity of the response task (Gower and Zylstra 1990).

Highlights from two of the studies, the Census of the Construction Industry and the National Training Survey, are discussed below.

5.1 Census of the Construction Industry

The annual Census of the Construction Industry was designed to provide comprehensive statistics on the construction industry in Canada. The target population consisted of establishments whose main revenue was derived from construction activity. There were two separate questionnaires for (a) General Contractors and Developers and (b) Trade Contractors and Sub-Contractors. The questionnaires, which were mailed to respondents, collected data on revenues and costs, labour data, and output distributions.

The questionnaires used in 1988 for the Census of the Construction Industry were redesigned for the 1989 survey. The main objectives of the revision were to reduce the content and response burden and to respond to the need for major improvements to the existing questionnaires.

A pretest of the revised questionnaires took place to obtain the reactions of contractors (Statistics Canada 1989). The pretest indicated that the revised forms were well received and understood by respondents. Some areas for further improvement such as changes to question wording and the clarification of certain instructions were identified.

To learn more about how respondents would view the revised questionnaires and to ensure that response rates and data quality would be maximized, further testing of the questionnaires using focus groups and cognitive methods was carried out in early 1990. This phase of testing was designed to obtain in-depth information on the following issues:

- How respondents felt about the questionnaires.
- The process that respondents went through to provide the information.
- The layout, presentation, and readability of the questionnaires.
- The extent to which respondents read and understood instructions and questions.
- Problems encountered by respondents while completing the questionnaires.
- Whether instructions and definitions were necessary, understandable, and useful.
- The accuracy of information provided by respondents.
- The use of estimates by respondents and their accuracy.
- The types of records from which information was obtained.
- The compatibility of the questions and response categories with respondents' record-keeping practices.
- Response burden in terms of time and effort.

The scope of the research included both the General Contractors and Developers questionnaire and the Trade Contractors and Sub-contractors questionnaire. Approximately 50 construction firms participated in the study. They were chosen to represent the types of respondents who completed the Census of the Construction Industry questionnaires. Twenty-five in-depth interviews, 16 concurrent think-aloud interviews, and 2 focus groups were conducted in Ottawa, Montréal and Toronto. All one-on-one interviews took place at the respondent's place of business.

A very interesting finding from the study was that there were two distinct groups of respondents. The first group of respondents included the president or vice-president of a company, who often had to consult other individuals to complete certain questions. It took these participants 35 to 45 minutes to complete the questionnaire. They were more likely to make estimates based on their familiarity with the company and were less concerned about accounting for differences between the questionnaire and the source of information used to complete the form.

On the other hand, respondents such as office managers, accountants and comptrollers took 75 to 90 minutes to complete the questionnaire. These respondents were much more concerned with detail and providing accurate answers. They were more likely to use multiple sources of information and to make calculations in answering the survey questions (Gower and Zylstra 1990; Gower and Nargundkar 1991).

Many respondents indicated that completing the questionnaire was not a priority. They viewed the survey as only one of the many forms and questionnaires that they had to complete each year. Many participants indicated that they often waited for the follow-up telephone call, and some even preferred, to answer the questionnaire over the telephone. They said that, over the telephone, they could

make estimates "off the tops of their heads" instead of carefully completing the form, and this required much less time and effort on their part.

The response burden was more perceived than real. Upon completing the questionnaire, many respondents remarked that it took surprisingly less time and was easier to complete than they had anticipated.

A common theme that emerged during the interviews and focus groups was the perceived value of the information being collected. Respondents wanted to know the purpose of completing the questionnaire and often questioned the value of the information to themselves and to other users of the information. Therefore, a major finding of the research was that the value of providing the information must be made clear to respondents. They wanted to know how the survey results were going to be used. They were also interested in learning how they could access the data.

Overall, the questionnaires were very well received by respondents. They appreciated the "business-like" appearance and approach of the questionnaires. Many were familiar with completing previous questionnaires for the Census of the Construction Industry. They felt that the redesigned forms were an improvement over the previous versions because they seemed shorter and less complicated. This was positive feedback and reassurance for the survey managers who designed the new questionnaires (Gower and Zylstra 1990; Price Waterhouse Management Consultants 1990).

The study identified many specific findings about how the questionnaires could be improved and made more "respondent-friendly". While the pretest provided valuable feedback about response rates and the completeness of reporting, the focus groups and cognitive research added significantly to these findings by providing in-depth, first-hand information about *how* and *why* respondents reacted to the questions as well as about *how* and *why* responses were chosen.

Figures 1 and 2 illustrate a few of the specific findings and how the questionnaire was improved based on these findings (Gower 1993). Figure 1 shows parts of Sections 2 and 4 of the 1988 version of the questionnaire for General Contractors and Developers, *before testing*. Figure 2 shows the corresponding parts of the final version of this questionnaire, *after testing*.

Section 2 – Statement of Income

On the final version of the questionnaire (Figure 2):

- A statement is provided at the beginning of Section 2, telling respondents that they could include their company's Financial Statements. On the version of the form (Figure 1) that was tested, many respondents missed this instruction because it appeared on a separate page of instructions.

Figure 1 (before testing): 1988 Census of the Construction Industry (General Contractors and Developers), Statistics Canada

SECTION 2. STATEMENT OF INCOME				Dollars (Omit cents)	
REVENUE				101	
2.1 Revenue from construction contracts .					
2.2 Other operating revenue, please specify:					
Type		Value			
102		103	\$		
104		105			
106		107			
108		109			
Total				110	
2.3 Total gross operating revenue (sum of items 2.1 and 2.2) . .				111	
2.4 Accounting method is: 1 <input type="checkbox"/> completed contract					
2 <input type="checkbox"/> percentage of completion					
DIRECT COST					
2.5 Work in progress, opening (add, if required for direct cost calculation)				112	
If direct cost detail is not available, please report percentages of total (item 2.15, sum should equal 100).					
2.6 Sub-contracts .				113	
2.7 Materials and supplies used (adjusted for change in inventory)				114	
2.8 Wages paid to hourly-rated employees (gross, before deductions for income tax, pension plans, insurance, etc.)				115	
2.9 Direct salaries paid to site supervisors, etc. (gross, before deductions for income tax, pension plans, insurance, etc.)				116	
2.10 Employee benefits (employer contributions not included in 2.8 and 2.9, such as pension plans, insurance, etc.)				117	
2.11 Land				118	
1 <input type="checkbox"/> undeveloped land					
Cost includes (please check): 2 <input type="checkbox"/> services, carrying charges, etc.					
3 <input type="checkbox"/> serviced lots					
2.12 Repair and maintenance of machinery and equipment .				119	
2.13 Equipment rental (without operator)				120	
2.14 Other direct cost . .				121	
2.15 Total direct cost (sum of items 2.6 to 2.14)				122	
2.16 Work in progress, closing (deduct if required for direct cost calculation)				123	
2.17 Total direct cost charged to contracts (item 2.5 plus 2.15 minus 2.16)				124	

SECTION 4. LABOUR FORCE			
4.1 For wages paid to your hourly paid labour force, reported in item 2.8, please report hours worked:			
201		hrs.	or average hourly rate: \$ 202 / hour
N.B.: Reported figure should be hours worked, i.e. one hour overtime paid at time and a half should be counted as one hour.			
4.2 For direct salaries paid, reported in item 2.9 please provide average annual number of employees:			
203		employees	
4.3 For overhead salaries paid, reported in item 2.19 please provide average annual number of employees:			
204		employees	

Figure 2 (after testing): 1989 Survey of the Construction Industry (General Contractors and Developers), Statistics Canada

SECTION 2. STATEMENT OF INCOME 201			Dollars (Omit cents)	
Instead of completing this section, you may include your company's Financial Statements, together with your otherwise completed questionnaire. If financial statements are included, go directly to Section 3.				
REVENUE				
2.1 Revenue from construction contracts .			202	
2.2 Other operating revenue, such as sales of materials, land sales, project or construction management, rentals of equipment and buildings, snow removal, consulting engineering fees. <i>Please specify:</i>				
Description				
			203	
			204	
			205	
			206	
2.3 Total gross operating revenue (sum of items 202 and 207-210)			211	
2.4 Please check accounting method used: 1 <input type="checkbox"/> complete contract				
2 <input type="checkbox"/> percentage of completion 212				
DIRECT COSTS				
2.5 Work in progress, opening (<i>add, if required for direct cost calculation</i>). Work in progress is defined as inventory of uncompleted and unbilled construction work done			213	
Only if direct costs detail is not available, please estimate percentages of total direct costs (<i>item 234, sum should equal 100</i>)				
2.6 Sub-contracts (<i>include equipment rental with operator</i>)			214	
2.7 Equipment rental without operator			215	
2.8 Materials and supplies used (<i>adjusted for change in inventory</i>)			216	
2.9 Wages paid to any hourly-rated employees (<i>gross, before deductions for income tax, pension plans, insurance, etc.</i>)			217	
2.10 Direct salaries charged to contract and paid to permanent staff, such as foremen, site supervisors, etc. (<i>gross, before deductions for income tax, pension plans, insurance, etc.</i>) . . .			218	
2.11 Employer portion of employee benefits, such as pension plans and insurance. (<i>Report only if employee benefits are not included in wages and direct salaries above</i>)			219	
2.12 Cost of land included in sales			220	
2.13 Repair and maintenance of machinery and equipment .			221	
2.14 Depreciation charged to contracts . .			222	
2.15 Other direct costs (<i>any other direct costs not separately reported above, such as pre-construction costs, site costs, fees, advertising, fuel, etc.</i>)			223	
2.16 Total direct cost (sum of items 224 to 233)			234	
2.17 Work in progress, closing (<i>deduct if required for direct cost calculation</i>) For definition of work in progress see question 2.5 above .			235	
2.18 Total direct costs charged to contract (<i>item 213 plus 234 minus 235</i>) .			236	

SECTION 4. LABOUR FORCE		4.2 Please report the average annual number of direct salaried employees (<i>whose salaries were reported in item 228</i>):	
4.1 Please report hours worked by your hourly paid labour force (<i>whose wages were reported in item 227</i>): N.B.: Reported figure should be hours worked, i.e. one hour overtime paid at time and a half should be counted as one hour. Figures for hours worked may be obtained from payroll records or Workers Compensation Board reports.		<div style="border: 1px solid black; padding: 2px; display: inline-block;">403</div> <div style="border: 1px solid black; width: 100px; height: 20px; display: inline-block;"></div> employees <i>Exclude owners and partners of unincorporated businesses</i>	
<div style="border: 1px solid black; padding: 2px; display: inline-block;">401</div> <div style="border: 1px solid black; width: 100px; height: 20px; display: inline-block;"></div> hours		4.3 Please report the average annual number of overhead salaried employees (<i>whose salaries were reported in item 237</i>):	
Only if hours worked are not available, please report average (straight-time) hourly rate:		<div style="border: 1px solid black; padding: 2px; display: inline-block;">404</div> <div style="border: 1px solid black; width: 100px; height: 20px; display: inline-block;"></div> employees <i>Exclude owners and partners of unincorporated businesses</i>	
<div style="border: 1px solid black; padding: 2px; display: inline-block;">402</div> \$ <div style="border: 1px solid black; width: 100px; height: 20px; display: inline-block;"></div> / hour		4.4 Number of professional engineers included in item 404:	
		<div style="border: 1px solid black; padding: 2px; display: inline-block;">405</div> <div style="border: 1px solid black; width: 100px; height: 20px; display: inline-block;"></div> engineers	

- Reference is made to line numbers (e.g., 202 and 207-210) instead of item numbers (e.g., 2.1 and 2.2). Although the line numbers are actually data code numbers, respondents viewed them as line numbers because they appeared similar to the common and well-known use of line numbers on the Canadian Income Tax forms.
- Important information such as definitions and what to include are provided in the items themselves instead of on the Instructions page.
- Respondents are only required to report estimated percentages if detail about direct costs is not available. This choice has been made clearer by printing “or” in large and bold print.

Note that, in completing Section 2, respondents consulted the following types of records: financial statements, on-line accounting systems, progress or work-on-hand billings, project reports, general ledgers, working papers, and audit statements.

Section 4 – Labour Force

On the final version of the questionnaire (Figure 2):

- Question 4.1 includes information that “hours worked” may be obtained from “payroll records or Workers’ Compensation Board reports”. During the think-aloud interviews, respondents noted that they consulted these types of records for the information.
- Clarification is provided that “average hourly rate” is to be reported “only if hours worked are not available”.
- Important information and instructions are included in the question items. For example, during testing, most respondents did not exclude owners and partners in reporting the numbers of employees in items 4.2 and 4.3 (even though this was specified on the Instructions page).

5.2 National Training Survey (NTS)

Two separate research studies, each involving the application of focus groups and cognitive research methods, have been used during the development and testing of the questionnaire for the National Training Survey (NTS).

The purpose of the NTS is to collect information on employee training and development in the private business sector. Respondents are asked to provide data on the type and volume of training, the number of trainees and their occupational groupings, the characteristics of the businesses providing training to their employees, and the amount of money being spent on this activity. In large businesses, respondents are the persons involved in the human resource planning and training areas of their company, while in smaller businesses they are typically the owner or chief executive officer.

At an early stage in developing the questionnaire, focus groups and in-depth interviews were held with representatives from small, medium and large companies. These methods were used because Statistics Canada felt it was

important to consult representatives of the business community to ensure that their interests and concerns about training were considered in the design of the NTS questionnaire.

The focus groups and interviews evaluated the clarity and appropriateness of terminology and concepts associated with the training of employees within a business establishment. The study investigated respondents’ understanding of terms such as “formal training” and “informal training” as well as their ability to use these terms to categorize their training activities.

Findings from this early phase of testing illustrated the importance of consulting with respondents before finalizing the terminology and concepts used in questionnaires. The findings from the study provided the survey project team with important information and insights into how the survey questions should be worded and how response options should be categorized.

For example, a significant finding from the focus groups and in-depth interviews was that many companies did not use the terms “formal” or “informal” to describe training activities and did not see the advantage or need to differentiate between the two terms. Many also perceived that there was no clear distinction between the terms “formal” and “informal” that would enable easy categorization of training activities.

The study helped the survey designers understand how respondents interpret terms and concepts. Participants provided suggestions on the appropriate terminology for them. For example, although they had difficulties with the terms “formal” and “informal,” participants were able to provide characteristics to define these terms. They described formal training as having “a formal structured curriculum or course outline with a beginning, middle and an end; that it has known objectives or clearly defined goals; that it has an evaluation component; . . . [and] that [it] has a dollar cost.” On the other hand, most participants perceived “informal training” to be on-the-job training having no structure, often involving learning by observing. “Lack of evaluation” was another characteristic often suggested to define informal training.

Another interesting finding was that many participants made a distinction between “training” and “developmental or educational activities”. The term “training” was not seen to cover all the activities that employers provide to support employee development. Some participants viewed “training” as job-specific and related to job productivity, and “development” as related to increasing the knowledge base of the individual (Kennedy and de Groh 1992).

After the draft NTS questionnaire was developed, it was tested using focus groups and concurrent think-aloud interviews. Representatives of a variety of businesses as well as a mixture of small, medium and large firms participated in the study. The study examined the following issues:

- The most appropriate person within a business to respond to the survey.
- How best to reach respondents.
- The process that respondents went through to provide the information.
- The way in which respondents understood the questions and instructions.
- Respondents' reaction to vocabulary and the groupings and classifications of occupations in the survey.
- Whether the information sought in the survey was readily available.
- The types of records from which information was obtained.
- The compatibility of the questions and response categories with respondents' record-keeping practices.
- Whether the reference periods requested in the survey corresponded to the record-keeping practices of respondents.
- Response burden in terms of time and effort.

Seven focus groups and 26 interviews were conducted in Ottawa, Toronto, Montréal, and Vancouver. In the final report (D.R. Harley Consultants Limited 1993), the Contractor reported many findings and made several recommendations to improve the questionnaire.

As in other studies of business surveys, a major finding was that many participants questioned the purpose behind the survey. They wanted to know why the information was being collected and how the survey results were going to be used. A strong theme that emerged throughout the focus groups and interviews was that respondents wanted to know "What's in this for me?"

Some participants suggested that the data be aggregated nationally, provincially and by sector so that they could compare themselves to other companies in their areas of business and in their part of the country. As one respondent said, "I would want the data to be specific to our industry with the volume and type of training that's being provided . . . It should allow us to compare ourselves to others in our sector – number of employees being trained and the percentage of payroll being spent on employee training."

Many small and medium-sized business respondents found the questionnaire too broad and the level of detail too complicated for them to answer. In their opinion, the questionnaire was designed for larger organizations. For example, many small businesses felt that they could not fit themselves into the categories provided by the questionnaire. They felt that much of their training fell into the "unstructured" category, and that the questionnaire was not capturing this aspect of training. However, at the same time, there were other respondents from small and medium-size businesses who commented that the questionnaire was thorough and complete.

The larger businesses also had difficulty with the level of detail being requested by the survey. The major problem

was that they keep training records by type of training that employees receive rather than by the occupational category of the people being trained.

Overall, a variety of record-keeping practices were observed. Some businesses keep excellent records on training, while others do not. Participants, who did not keep good records or whose records did not contain the requested information, found the questionnaire difficult to answer. Others, who had sophisticated records, could manipulate their data to fit the questionnaire. The one exception was the questions on training expenditure for which they found it difficult to provide detailed information. Global figures were more easily available, they said. Many businesses indicated that their training records were not centralized, thus making the questionnaire more difficult and requiring longer time to complete. They said that they would complete what they could, and then coordinate the completion of the rest of the questionnaire by forwarding it to many parts of their organization.

Although many participants were initially overwhelmed by the size and apparent complexity of the questionnaire, they found it easier to complete than expected. Many found that the thoroughness of the questionnaire actually made them remember many training activities that they would not ordinarily have reported on.

Most participants felt that the questionnaire should be shorter. But they also suggested adding a few more open-ended questions about future training. In terms of response burden, respondents (especially in medium-sized and large-size companies) found that the questions about training expenses, training hours, and the numbers of employees trained by occupational categories would require hours of work to compile.

Differences were found in the time it took respondents to complete the questionnaire. Small businesses took between 10 minutes and 1 hour to complete the questionnaire. Large businesses, on the other hand, estimated that it would take about 2 hours to complete the questionnaire (D.R. Harley Consultants Limited 1993).

6. CONCLUDING REMARKS

This paper has provided an overview of questionnaire design for business surveys. As the paper has pointed out, many considerations go into designing business survey questionnaires. They include the survey's objectives and data requirements as well as consultation with data users and respondents on the nature and concerns of the respondent population. Other considerations are response burden, the method of data collection, the availability of data, and the use of records, as well as the need for testing the questionnaires.

Specific design issues that should be taken into account include the instructions, the clarity and readability of the

questions, the logical sequencing of the questions, the compatibility of response categories and reference periods with respondents' record-keeping practices, and data processing requirements. The questionnaire should be respondent-friendly and interviewer-friendly.

To ensure the collection of accurate and useful data in business surveys, it is important to understand the response process that respondents go through in completing a questionnaire. Focus groups and cognitive research methods are very effective ways to study this response process and to test questionnaires. They provide the opportunity to consult directly with respondents and, thereby, to bring their ideas, concerns, and suggestions into the questionnaire design process.

Looking towards the future, research and experience should lead to improvements in the methods and approaches that are currently used to develop and test business survey questionnaires. An important area that requires more research and development is the relationship among the questionnaire, the respondent, and the external information source as well as the influence that this relationship has on the response process and the accuracy of reporting.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the work of the following consultants and contractors in undertaking this research: D.R. Harley Consultants Limited, Kennedy and de Groh Consultants, and Price Waterhouse Management Consultants. The views expressed are those of the author, and do not necessarily reflect those of Statistics Canada nor these contractors. The author also wishes to thank the referee for helpful comments.

REFERENCES

- BUREAU, M. (1991). Experience with the use of cognitive methods in designing business survey questionnaires. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-717.
- DEMAIO, T.J. (Ed.) (1983). *Approaches to Developing Questionnaires*. Statistical Policy Working Paper 10, Washington, DC: United States Office of Management and Budget.
- EDWARDS, W.S., and CANTOR, D. (1991). Toward a response model in establishment surveys. In *Measurement Errors in Surveys*. (Eds. Paul P. Biemer *et al.*). New York: John Wiley and Sons, 211-233.
- GOSS, GILROY and ASSOCIATES LTD. (1989). Qualitative Research to Evaluate the Questionnaire of the Survey of Employment, Payrolls and Hours (SEPH). Final report submitted to Statistics Canada.
- GOSS, GILROY and ASSOCIATES LTD. (1990). Qualitative Research to Evaluate the Redesigned Survey Materials of the Survey of Employment, Payrolls and Hours (SEPH). Final report submitted to Statistics Canada.
- GOWER, A.R. (1993). Questionnaire design for establishment surveys. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 950-956.
- GOWER, A.R. (1991). The Questionnaire Design Resource Centre's Role in Questionnaire Research and Development at Statistics Canada. 48th Session of the International Statistical Institute. *Booklet*, Volume III, 58-59.
- GOWER, A.R., and NARGUNDKAR, M.S. (1991). Cognitive Aspects of Questionnaire Design: Business Surveys versus Household Surveys. *Proceedings of the 1991 Annual Research Conference*. Washington, DC: United States Bureau of the Census, 299-312.
- GOWER, A.R., and ZYLSTRA, P.D. (1990). The Use of Qualitative Methods in the Design of a Business Survey Questionnaire. Contributed Paper (unpublished). International Conference on Measurement Errors in Surveys, Tucson, Arizona.
- D.R. HARLEY CONSULTANTS LIMITED (1993). Qualitative Testing of the Draft National Training Survey Questionnaire. Final report submitted to Statistics Canada.
- KENNEDY and DE GROH CONSULTANTS (1992). Testing of Definitions for the National Training Survey. Final report submitted to Statistics Canada.
- NOONAN, M. (1992). Final report on Personal Interviews with Potential Respondents for the Proposed Wholesale and Retail Trades Survey. Unpublished Report, Statistics Canada.
- PRICE WATERHOUSE MANAGEMENT CONSULTANTS (1990). Qualitative Research Related to the Re-design of the Census of the Construction Industry Questionnaires. Final report submitted to Statistics Canada.
- STATISTICS CANADA (1989). Construction Census Questionnaire Test. Unpublished Report, Construction Census Section, Industry Division.
- STATISTICS CANADA (1994). Policy on the Development, Testing and Evaluation of Questionnaires.
- STATISTICS CANADA (1986). Policy on Informing Survey Respondents.
- TOURANGEAU, R. (1984). Cognitive sciences and survey methods. In *Cognitive Aspects of Survey Methodology: Building a Gap Between Disciplines*. (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau). Washington, DC: National Academy Press, 73-100.

Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse

E. RANCOURT, H. LEE and C.-E. SÄRNDAL¹

ABSTRACT

Most surveys suffer from the problem of missing data caused by nonresponse. To deal with this problem, imputation is often used to create a "completed data set", that is, a data set composed of actual observations (for the respondents) and imputations (for the nonrespondents). Usually, imputation is carried out under the assumption of unconfounded response mechanism. When this assumption does not hold, a bias is introduced in the standard estimator of the population mean calculated from the completed data set. In this paper, we pursue the idea of using simple correction factors for the bias problem in the case that ratio imputation is used. The effectiveness of the correction factors is studied by Monte Carlo simulation using artificially generated data sets representing various super-populations, nonresponse rates, nonresponse mechanisms, and correlations between the variable of interest and the auxiliary variable. These correction factors are found to be effective especially when the population follows the model underlying ratio imputation. An option for estimating the variance of the corrected point estimates is also discussed.

KEY WORDS: Conditional bias; Monte Carlo simulation; Restoring estimator; Variance estimation.

1. INTRODUCTION

Occurrence of nonresponse is rather a norm than an exception in surveys. Missing data caused by nonresponse are often imputed to obtain a completed data set and the standard estimator is applied to the completed data set assuming that the underlying response mechanism is unconfounded. However, a point estimate obtained in such a way is biased when the response mechanism is confounded. The bias in this case could be very severe as pointed out in Lee, Rancourt and Särndal (1994). A response mechanism is unconfounded, according to Rubin (1987, p. 39), if it does not depend on the variable under study, otherwise it is confounded. (A formal definition suitable for this paper will be given in Section 2.)

In a Bayesian framework, a concept similar to that of an unconfounded response mechanism is termed ignorable. For bias caused by a nonignorable response mechanism, Rubin (1977, 1987) and Little and Rubin (1987) considered a method to correct the respondent mean using auxiliary variables. In this approach, a linear regression is assumed between the variable of interest y and a vector of auxiliary variables x . The regression coefficient vector for the nonrespondents is assumed to have a normal prior with mean equal to the regression coefficient vector for the respondents.

Assuming a logistic model for the response probability, Greenless, Reece and Zieschang (1982) proposed a method to deal with nonignorable nonresponse using maximum likelihood estimation. Further, a linear regression model is assumed for the relationship between y and x , a vector

of auxiliary variables. The logistic model of the response probability includes y and z , a vector of other auxiliary variables. Assuming also that the error term of the regression is normally distributed, they obtain maximum likelihood estimates of the unknown parameters of the regression model and the logistic model. Finally, for a nonrespondent, an imputed value is calculated as the mean of the distribution of y conditional on the values of x and z for the nonrespondents, and the estimated parameters. Such a method may give good results when all the model assumptions are satisfied but is likely to be highly sensitive to the specifications of the two models. The adequacy of the response probability model is usually untestable. If data are available from an external source, however, then it may be possible to test the response probability model as Greenless *et al.* did in their application to the Current Population Survey data. This method is highly computer-intensive.

In the case of categorical data, a few methods have also been proposed to deal with the problem of nonignorable nonresponse. For instance, Baker and Laird (1988) try to model the response mechanism with the help of log-linear models. As well, causal modeling is discussed in Fay (1986, 1989).

Ratio imputation is often used at Statistics Canada, especially in repeated surveys. For instance, in the Monthly Survey of Manufacturing, a missing value of the current shipment is imputed by ratio imputation using previous month shipment as the auxiliary variable value. This simple method is very appealing to subject matter specialists because it reflects month-to-month movement.

¹ E. Rancourt and H. Lee, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; C.-E. Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec), Canada, H3C 3J7.

In this paper, we investigate the possibility of improving the estimator applied to data containing ratio imputation with the aid of simple correction factors. Therefore, we assume that imputation has already been performed, and try to correct the estimator. We focus our attention on the estimation of the mean. The use of simple correction factors would be very appealing to the user provided it works reasonably well. Such a procedure is also easy to implement without resorting to excessive computational efforts and it enables us to avoid explicit modeling of the nonresponse mechanism. However, our approach differs from Rubin's in that we use sample dependent correction factors rather than an *a priori* chosen constant.

In Section 2, we define several simple correction factors that meet our requirements. In Section 3, we propose a variance estimator that may be used in conjunction with the corrected point estimators. The properties of the corrected point estimators were examined by a Monte Carlo simulation reported in Sections 4 and 5. Section 6 presents some concluding remarks.

2. SIMPLE BIAS CORRECTION FACTORS

Let $U = \{1, \dots, k, \dots, N\}$ denote the index set of a finite population and let the population mean of the variable of interest y be denoted by $\bar{y}_U = (1/N) \sum_U y_k$. We assume that $y_k > 0$ for all $k \in U$. From U , a simple random sample s of size n is drawn without replacement (SRSWOR). The unbiased estimator that would be used with 100% response is the sample mean

$$\bar{y}_s = (1/n) \sum_s y_k. \quad (2.1)$$

Let r and o be the sets of the responding and non-responding units, respectively, so that $s = r \cup o$. We denote the SRSWOR sampling plan by $p(\cdot)$ and the response mechanism given s by $q(\cdot | s)$. That is, $p(s)$ is the probability that the SRSWOR sample s is drawn, and $q(r | s)$ is the probability that the set r responds given the sample s . Let also m and l be the sizes of r and o , respectively. For simplicity, we assume that the probability of $m = 0$ is negligible. We assume that imputation is carried out with the aid of an auxiliary variable, x , whose value, x_k , is known and positive for all $k \in s$. If $k \in o$, the missing value y_k is imputed by \hat{y}_k . The completed data set is denoted as $\{y_{\cdot k} : k \in s\}$ where $y_{\cdot k} = y_k$ if $k \in r$ and $y_{\cdot k} = \hat{y}_k$ if $k \in o$.

In this paper, we examine ratio imputation. This often-used imputation method is based on a simple model. That is, if the value y_k is missing, it is imputed by $\hat{B}_r x_k$, where $\hat{B}_r = (\sum_r y_k) / (\sum_r x_k)$. The model denoted ξ , is stating that, for $k \in s$,

$$y_k = \beta x_k + \epsilon_k, \quad E_\xi(\epsilon_k | x_k) = 0, \quad V_\xi(\epsilon_k | x_k) = \sigma^2 x_k, \\ E_\xi(\epsilon_k \epsilon_l | x_k, x_l) = 0, \quad k \neq l. \quad (2.2)$$

Under this model, $\hat{B}_r x_k$ is the best linear unbiased predictor of the missing value y_k , based on the respondent data $\{(y_k, x_k) : k \in r\}$. The completed data set is then composed of the values

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}_r x_k, & \text{if } k \in o. \end{cases} \quad (2.3)$$

The customary procedure is to apply the estimator formula used for 100% response to the completed data set. This gives

$$\bar{y}_{\cdot s} = \frac{1}{n} \sum_s y_{\cdot k} = \frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s = \bar{y}_{\text{raimp}}, \quad (2.4)$$

where $\bar{x}_s = (1/n) \sum_s x_k$, $\bar{y}_r = (1/m) \sum_r y_k$ and $\bar{x}_r = (1/m) \sum_r x_k$. Note that *raimp* stands for ratio imputed.

It now becomes necessary to address the question whether the imputation can restore the full response estimator, \bar{y}_s , in the sense that the imputation estimator $\bar{y}_{\cdot s}$ is equal to \bar{y}_s in expectation given s . Unless this can be achieved, the ratio imputation will have introduced bias. To examine this question, we must consider the response mechanism. A response mechanism $q(\cdot | s)$ is said to be *unconfounded* for the purpose of this paper if it is of the form $q(r | s) = q(r | x_s)$, where $x_s = \{x_k : k \in s\}$ and the response probabilities satisfy $P(k \in r | s) > 0$ for all $k \in s$. That is, it may depend on s and on the associated x -values. If it depends also on the y -values, so that $q(r | s) = q(r | x_s, y_s)$, then it is called *confounded*. In these definitions, the response mechanism is conditional on the realized sample s . Slightly different definitions of “confounded” and “unconfounded” are given in Rubin (1987, p. 39) where they are unconditional.

An example of an unconfounded response mechanism is

$$q(r | s) = \prod_{k \in r} (1 - \Theta_k) \prod_{k \in s-r} \Theta_k,$$

where $\Theta_k = 1 - P(k \in r | s) = 1 - e^{-\gamma x_k}$ for some positive constant γ , is the nonresponse probability of unit k . By contrast, if $\Theta_k = 1 - e^{-\gamma y_k}$, then $q(r | s)$ is a confounded mechanism.

A particularly simple unconfounded mechanism is the uniform response mechanism defined by $q(r | s) = (1 - \Theta)^m \Theta^{n-m}$. Here, units respond according to independent and identical Bernoulli $(1 - \Theta)$ trials, where Θ is the nonresponse probability common to all units.

Whether an imputation estimator \hat{y}_U of \bar{y}_U , including \bar{y}_{raimp} given by (2.4), is considered good depends in part on the assumptions made by the analyst about the response mechanism and in part on the relation between y and x . Several possible assumptions are discussed later in this section. For any given s , the goal is that, under specified realistic assumptions, the expectation of the difference

$\hat{y}_U - \bar{y}_s$ should be close to zero. That is, under the given assumptions, the conditional bias of \hat{y}_U , $C\text{-bias}(\hat{y}_U) = E(\hat{y}_U - \bar{y}_s | s)$, should be small. We call \hat{y}_U a *restoring estimator* of \bar{y}_U if $C\text{-bias}(\hat{y}_U) = 0$ or ≈ 0 , that is, if \hat{y}_U is (approximately) equal to \bar{y}_s in conditional expectation. It follows that if the $C\text{-bias}$ is (approximately) zero for any s , then the unconditional bias over all sample realizations s is also (approximately) zero.

Different analysts make different assumptions. Let us consider some typical assumptions and ask the question: What restoring estimators do these assumptions allow?

Assumption I: The response mechanism is uniform.

Under Assumption I, \bar{y}_{raimp} is a restoring estimator. To see this, note that

$$C\text{-bias}(\bar{y}_{\text{raimp}}) = E_q(\bar{y}_{\text{raimp}} | s) - \bar{y}_s \approx 0,$$

because, given s , \bar{y}_{raimp} is the classical ratio estimator of \bar{y}_s . Assumption I is unrealistic in most surveys. The response propensity is known to vary with observable characteristics such as size and industry (for business establishments), family size and type (for households), age, sex and income (for individuals). Under this unrealistic assumption, even a naive estimator such as the respondent mean, $\bar{y}_r = (1/m) \sum_r y_k$, is a restoring estimator:

$$C\text{-bias}(\bar{y}_r) = E_q(\bar{y}_r | s) - \bar{y}_s = 0.$$

However, if Assumption I holds, \bar{y}_{raimp} is preferred to \bar{y}_r because the ratio estimator feature leads to a smaller variance if the model ξ holds.

The analyst clearly needs to consider more realistic assumptions which allow the response probabilities to vary with background variables. The following assumption, composed of two parts, is of this kind.

Assumption II: (II-1): the response mechanism is unconfounded but otherwise arbitrary;

(II-2): the ratio model (2.2) holds.

Here (II-1) is a weaker and more realistic requirement on the response mechanism than the uniformity requirement in Assumption I. Under (II-1), the response mechanism can be of any form as long as it is unconfounded. However, Assumptions I and II are not directly comparable since II contains a model component, (II-2), which is lacking in I. Under Assumption II, \bar{y}_{raimp} is a restoring estimator because

$$\begin{aligned} C\text{-bias}(\bar{y}_{\text{raimp}}) &= E_\xi\{E_q(\bar{y}_{\text{raimp}}) - \bar{y}_s | s\} \\ &= E_q E_\xi\left(\frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s\right) - E_\xi(\bar{y}_s) \\ &= E_q(\beta \bar{x}_s) - \beta \bar{x}_s = 0. \end{aligned}$$

Note that changing the order of the expectations, $E_\xi E_q$ to $E_q E_\xi$, is allowed under Assumption II, because the response mechanism is then of the form $q(r | x_r)$, that is, it does not depend on the y -values. By contrast, the respondent mean \bar{y}_r is not a restoring estimator because

$$C\text{-bias}(\bar{y}_r) = E_\xi\{E_q(\bar{y}_r) - \bar{y}_s | s\} = \beta\{E_q(\bar{x}_r | s) - \bar{x}_s\},$$

which is generally nonzero under Assumption II. We can, however, transform \bar{y}_r into a restoring estimator by the use of a multiplicative correction factor. This leads to

$$\bar{y}_r \left\{ 1 + \left(1 - \frac{m}{n} \right) \left(\frac{\bar{x}_o}{\bar{x}_r} - 1 \right) \right\}, \quad (2.5)$$

which is just another way of writing \bar{y}_{raimp} , as can easily be verified. In an example using the Bayesian approach, Little and Rubin (1987, p. 233) arrive at an estimator identical to the estimator (2.5).

Let us now consider confounded response mechanisms. They cause more difficult problems for finding a restoring estimator.

Assumption III: (III-1): the response mechanism is confounded but otherwise arbitrary;

(III-2): the ratio model (2.2) holds.

It is usually difficult, if not impossible, for the analyst to decide whether Assumption II or Assumption III is more appropriate. Examining the data will not be of much help if the only data available relate to the present point in time, as would typically be the case in a one-time survey. The assumption made (whether II or III) is then unverifiable. By contrast, if the analyst has experience with a regularly repeated survey, he or she may have legitimate reasons to believe, for example, that the nonresponse is a function of the variable of interest.

In some situations, the assumption of a confounded mechanism may be made on the following grounds. Suppose in a survey of personal finances that y , the variable under study is "savings" and that x , the auxiliary variable is "income", with values x_k known for the individuals $k \in s$. The nonresponse probability of respondent k is likely to be correlated with the savings figure y_k that he or she is asked to reveal as well as with the income figure x_k known from other sources. But since savings, not income, is the variable with which the respondent is directly confronted in the survey, the assumption that the nonresponse probability is a function of y_k may be more realistic than the assumption that it is a function of x_k . Hence a confounded mechanism may be more realistic to assume than an unconfounded mechanism.

Under Assumption III, neither \bar{y}_r nor \bar{y}_{raimp} are restoring estimators. The $C\text{-bias}$ of \bar{y}_{raimp} can be expressed as

$$C\text{-bias}(\bar{y}_{\text{raimp}}) = \bar{x}_s E_\xi E_q \left(\frac{\sum_r \epsilon_k}{\sum_r x_k} \right),$$

where ϵ_k is defined by the model (2.2). This C -bias is generally nonzero and can be quite large when the non-response rate is high and the correlation is not so strong. However, the C -bias is hard to evaluate, since the exact form of the response mechanism is left unspecified. Note that changing the order of the expectations E_ξ and E_q is not permitted under Assumption III since $q(r | s)$ depends on the y -values. For example, a negative C -bias is likely to occur if the respondent residual total, $\sum_r \epsilon_k$ tends to be negative.

A confounded response mechanism (as in Assumption III), introduces bias in the slope estimator $\hat{B}_r = (\sum_r y_k) / (\sum_r x_k)$. Consequently, $\hat{B}_r x_k$ is a biased imputation for a missing value y_k . To improve the situation, suppose that a missing value y_k is imputed by $C \hat{B}_r x_k$ instead of $\hat{B}_r x_k$, where C is a quantity to be specified. Then the data after imputation are given by

$$y_{\cdot k}^c = \begin{cases} y_k, & \text{if } k \in r \\ C \hat{B}_r x_k, & \text{if } k \in o \end{cases} \quad (2.6)$$

and denoting the sample mean of these data as $\bar{y}_{\cdot s} = (1/n) \sum_s y_{\cdot k}^c$, we get the estimator

$$\bar{y}_{\cdot s} = \bar{y}_r \left[1 + \left(1 - \frac{m}{n} \right) \left(C \frac{\bar{x}_o}{\bar{x}_r} - 1 \right) \right]. \quad (2.7)$$

A simple correction of the type used in (2.6) was mentioned in Rubin (1986; 1987, p. 203) in the context of multiple imputation. Rubin views C as a fixed constant chosen by the user according to his or her prior knowledge. If such a choice happens to be well founded, the bias of (2.7) may be small.

Here, we shall examine choices of C that are adaptive, that is, they reflect the realized sample s and the realized response set r . Ideally, C should be such that the imputation will exactly restore the estimator $\bar{y}_s = (1/n) \sum_s y_k$ that would be used with 100% response. This C -value is determined by the equation

$$\bar{y}_s = \frac{1}{n} \sum_s y_k = \frac{1}{n} \sum_s y_{\cdot k}^c = \frac{1}{n} \left(\sum_r y_k + \sum_o C \hat{B}_r x_k \right).$$

A simple calculation shows that the optimal C -value is

$$C_{\text{opt}} = \frac{\hat{B}_o}{\hat{B}_r},$$

where $\hat{B}_o = \sum_o y_k / \sum_o x_k$ is the slope estimate if the model (2.2) could be fitted to nonrespondents. The imputed values would then be $\hat{y}_k = \hat{B}_o x_k$ for $k \in o$. Obviously, C_{opt} and \hat{B}_o cannot be computed since they depend on missing y_k -values. For an unconfounded mechanism (as in Assumption II), we can expect $C_{\text{opt}} \approx 1$, given s , because

$$E_\xi E_q (C_{\text{opt}} | s) = E_q E_\xi \left(\frac{\hat{B}_o}{\hat{B}_r} | s \right) \approx 1.$$

But for a confounded mechanism (as in Assumption III), C_{opt} can be distinctly away from unity. Suppose that $C_{\text{opt}} > 1$. Note that $C_{\text{opt}} > 1$ if and only if $\sum_r e_{ks} < 0$ with $e_{ks} = y_k - \hat{B}_s x_k$, where $\hat{B}_s = (\sum_s y_k) / (\sum_s x_k)$ is the unknown slope estimate with 100% response. That is, $C_{\text{opt}} > 1$ implies that respondents' residuals e_{ks} are negative on the average. An illustration of this is shown in figure 1, where $n = 10$, $l = n - m = 5$, and all five respondents' residuals e_{ks} are negative.

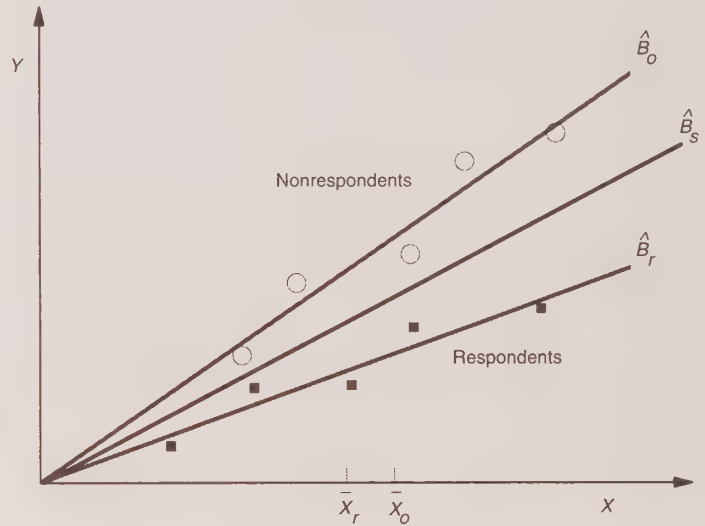


Figure 1. Example of data plot (y_k, x_k) for a confounded response mechanism.

Assuming that $C_{\text{opt}} > 1$, one approach for the analyst working under Assumption III is to choose a computable C likely to satisfy $C > 1$ and then use this C to construct the estimator (2.7). Factors C that will sometimes work in this manner are

$$c_1 = \frac{\bar{x}_o}{\bar{x}_r}, \quad c_2 = \frac{\bar{x}_o}{\bar{x}_s}, \quad c_3 = \frac{\bar{w}_o}{\bar{w}_r}, \quad c_4 = \frac{\bar{w}_o}{\bar{w}_s}. \quad (2.8)$$

They are based on the logic that if the response mechanism is confounded in such a way that the nonresponse probability is a function of y (for example, $\Theta_k = 1 - e^{-\gamma y_k}$

with $\gamma > 0$), then both $C_{\text{opt}} > 1$, and $\bar{x}_o > \bar{x}_r$ are likely to occur, as Figure 1 illustrates. Conversely, if nonresponse is a decreasing function of y_k , then both $C_{\text{opt}} < 1$, and $\bar{x}_o < \bar{x}_r$ are likely to occur.

One important feature of such correction factors is that they can, but need not, be calculated during the imputation phase. For instance, if the usual ratio imputation $\hat{B}_r x_k$ was carried out at the imputation phase, it is then possible to calculate a suitable correction factor at the estimation phase without changing the originally imputed values.

Note that c_2 implies a somewhat milder correction than c_1 : if $c_1 > 1$, we have $1 < c_2 < c_1$. The choices $C = c_3$ and $C = c_4$ are calculated on the ranks of the x -values, rather than on the x -values themselves, to dampen the effect of extreme x -values. More specifically, letting w_k be the rank of x_k in the data set $\{x_k : k \in s\}$, the w -means in c_3 and c_4 are $\bar{w}_s = (1/n) \sum_s w_k$, $\bar{w}_r = (1/m) \sum_r w_k$ and $\bar{w}_o = (1/l) \sum_o w_k$. The four estimators obtained by letting $C = c_i$ in (2.7) according to (2.8) will be denoted as $\bar{y}_{c_i \cdot s}$, $i = 1, \dots, 4$. In particular, we have

$$\bar{y}_{c_1 \cdot s} = \bar{y}_r \left[1 + \left(1 - \frac{m}{n} \right) \left\{ \left(\frac{\bar{x}_o}{\bar{x}_r} \right)^2 - 1 \right\} \right], \quad (2.9)$$

and

$$\bar{y}_{c_2 \cdot s} = \bar{y}_r \left[1 + \left(1 - \frac{m}{n} \right) \left\{ \frac{\bar{x}_o^2}{\bar{x}_r \bar{x}_s} - 1 \right\} \right]. \quad (2.10)$$

The correction factors given in (2.8) are not ideal when the correlation between x and y is close to 1. In this case, we have $\hat{B}_r \approx \hat{B}_s \approx \hat{B}_o$, provided that the model (2.2) holds. Therefore, the correction factor C should be close to 1. However, the correction factors given in (2.8) could be very different from 1 and using them would bring bias. For this reason, it may be preferable to work with a correction factor C in (2.7) that takes the correlation into account. Correction factors of this kind are

$$k_i = 1 - \{ (c_i^2 - 1) (\hat{R}_{xy}^2 - 1) \}, \quad (2.11)$$

where c_i , $i = 1, \dots, 4$, are the four correction factors given in (2.8), and \hat{R}_{xy} is the estimated correlation coefficient based on the respondent data. In our Monte Carlo simulation we also included the estimator (2.7) corresponding to the four choices $C = k_i$, $i = 1, \dots, 4$. These estimators will be denoted as $\bar{y}_{k_i \cdot s}$, $i = 1, \dots, 4$.

3. VARIANCE ESTIMATION

Since we are interested in variance estimators based on single value imputation, the variance estimation method proposed in Särndal (1990, 1992) is of interest. Assuming unconfounded nonresponse and that the model ξ in (2.3)

holds, the variance estimator for the point estimator \bar{y}_{raimp} in (2.4) obtained by this method is given by

$$\begin{aligned} \hat{V}(\bar{y}_{\text{raimp}}) &= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{\sum_s (y_{\cdot k} - \bar{y}_{\cdot s})^2}{n-1} \\ &\quad + \left(\frac{1}{n} - \frac{1}{N} \right) A_o \hat{\sigma}^2 + \left(\frac{1}{m} - \frac{1}{n} \right) A_1 \hat{\sigma}^2 \\ &= \hat{V}_{\text{ord}} + \hat{V}_{\text{dif}} + \hat{V}_{\text{imp}}, \end{aligned} \quad (3.1)$$

where

$$\begin{aligned} A_o &= \frac{1}{n-1} \left\{ \sum_o x_k - \frac{\sum_o x_k^2}{\sum_r x_k} + \frac{\bar{x}_s \sum_o x_k}{\sum_r x_k} \right\}, \\ A_1 &= \frac{\bar{x}_s \bar{x}_o}{\bar{x}_r}. \end{aligned}$$

and

$$\hat{\sigma}^2 = \frac{\sum_r e_k^2 / (m-1)}{\bar{x}_r \{ 1 - (\text{cv}_{xr})^2 / m \}}, \quad (3.2)$$

where

$$e_k = y_k - \hat{B}_r x_k, \quad \text{cv}_{xr} = \frac{\sqrt{\sum_r (x_k - \bar{x}_r)^2 / (m-1)}}{\bar{x}_r}.$$

The variance of \bar{y}_{raimp} has two components, namely, the sampling variance and the variance due to imputation. The first term in (3.1) (denoted by \hat{V}_{ord}) is an estimate of the sampling variance calculated using the ordinary variance formula assuming that imputed data are as good as real observations. Since this assumption does not hold, \hat{V}_{ord} underestimates the true sampling variance. To correct this underestimation, the second term \hat{V}_{dif} in (3.1) is added. The last term \hat{V}_{imp} in (3.1) is an estimate of the variance due to imputation.

If we compute the mean of the y -values from the completed data set $\{y_{\cdot k} : k \in s\}$ given in (2.6), we get the estimator (2.7). Its variance estimator should take the correction factor C into account. If we can assume that the expectation $E_{\xi} E_p E_q$ is equal to $E_p E_q E_{\xi}$ (this is true under unconfounded nonresponse), we can use Särndal's (1990, 1992) method to obtain a variance estimator which takes C into account. However, we are mainly interested in confounded cases. We are therefore proposing a variance estimator based on the following heuristic argument.

The estimator $\hat{\sigma}^2$ in (3.2) uses the respondent data only. It will certainly be biased for confounded mechanisms and some correction is needed in order to use formula (3.1) for the corrected estimator (2.7). We suggest to replace $\hat{\sigma}^2$ in (3.1) by $C^2\hat{\sigma}^2$, to obtain the following variance estimator for the estimator $\bar{y}_{c \cdot s}$ in (2.7):

$$\hat{V}(\bar{y}_{c \cdot s}) = \hat{V}_{\text{ord}}^c + C^2(\hat{V}_{\text{dif}} + \hat{V}_{\text{imp}}), \quad (3.3)$$

where \hat{V}_{ord}^c is computed using the data after imputation with the bias correction factor C . Replacing C^2 by c_i^2 or k_i^2 , we obtain the variance estimators corresponding to $\bar{y}_{ci \cdot s}$ or $\bar{y}_{ki \cdot s}$. The resulting variance estimators work quite well in many of the cases covered in the simulation reported in Section 5.

4. SIMULATION STUDY

We are considering eight corrected estimators corresponding to the eight correction factors given in (2.8) and (2.11). A simulation study was conducted to determine whether the corrected estimators succeed in restoring \bar{y}_s under different response mechanisms, in particular, confounded mechanisms. For comparison, we also included the uncorrected estimators \bar{y}_r and $\bar{y}_{\text{raimp}} = \bar{x}_s \bar{y}_r / \bar{x}_r$ given by (2.2). Our primary objective was to examine the corrected estimators when the finite population follows the ratio model ξ given by (2.3). However, we also wanted to see how the corrected estimators behave under relationships other than linear regression through the origin.

We also studied the coverage rates associated with the different estimators when the confidence intervals are computed with the aid of the variance estimators proposed in Section 3.

For the simulation, we generated 12 different finite populations, each of size $N = 100$, by specifying in different ways the constants a , b , c , and d in the regression model:

$$\begin{aligned} \mathbb{E}: y_k &= a + bx_k + cx_k^2 + \epsilon_k, \quad E_{\Xi}(\epsilon_k) = 0, \\ V_{\Xi}(\epsilon_k) &= d^2 x_k, \end{aligned} \quad (4.1)$$

where the ϵ_k are assumed to be independent. Four different regression types were created by four different specifications of (a, b, c) . These types are called RATIO ($a = c = 0, b > 0$, thus conforming to the ratio model ξ in (2.3)), CONCAVE ($a = 0, b > 0, c < 0$), CONVEX ($a = 0, b > 0, c > 0$) and NONRATIO ($a \neq 0, b > 0, c = 0$). For each regression type, three different levels of the model correlation ρ_{xy} , 0.7, 0.8 and 0.9, were obtained by a suitable choice of d . This resulted in 12 specifications of (a, b, c, d) as shown in Table 1.

Table 1
Characteristics of the Populations

POP	TYPE	a	b	c	d	R_{xy}	MEAN of y
1	RATIO	0	1.5	0	6.12	0.69	70.95
2	RATIO	0	1.5	0	4.50	0.81	69.92
3	RATIO	0	1.5	0	2.91	0.90	72.67
4	CONCAVE	0	3	-0.01	6.78	0.71	117.27
5	CONCAVE	0	3	-0.01	4.83	0.81	114.57
6	CONCAVE	0	3	-0.01	2.80	0.90	112.11
7	CONVEX	0	0.25	0.01	5.98	0.71	35.89
8	CONVEX	0	0.25	0.01	4.22	0.81	37.06
9	CONVEX	0	0.25	0.01	2.35	0.90	43.92
10	NON-RATIO	20	1.5	0	6.12	0.71	95.25
11	NON-RATIO	20	1.5	0	4.50	0.81	94.46
12	NON-RATIO	20	1.5	0	2.91	0.90	93.32

For each of the 12 specifications, we generated 100 population values (y_k, x_k) , $k = 1, \dots, 100$, by a two step process. We used the Γ -distribution with parameters α and β . Its density is

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) \quad \text{for } x > 0. \quad (4.2)$$

First, we generated 100 values x_k , $k = 1, \dots, 100$, according to the Γ -distribution with parameters $\alpha = 3$, $\beta = 16$, implying that the mean is $\alpha\beta = 48$ and the variance $\alpha\beta^2 = 768$. Then, for each fixed x_k , $k = 1, \dots, 100$, we generated one value y_k according to the Γ -distribution with parameters

$$\alpha = \frac{\{\mu(x)\}^2}{\sigma^2(x)} = \frac{(a + bx + cx^2)^2}{d^2 x}, \quad (4.3)$$

$$\beta = \frac{\sigma^2(x)}{\mu(x)} = \frac{d^2 x}{a + bx + cx^2}, \quad (4.4)$$

where $x = x_k$ and (a, b, c, d) is one of the 12 vectors fixed in advance. This implies that $E_{\Xi}(y_k | x_k) = \alpha\beta = a + bx_k + cx_k^2$ and $V_{\Xi}(y_k | x_k) = \alpha\beta^2 = d^2 x_k$, as required under the model (4.1). The same x -values were used for all 12 populations. For the populations generated by this process, Table 1 shows the values of the population correlation R_{xy} and the population mean of y . Note that the values of a , b , c , and d were chosen so as to obtain realistic types of populations that can be encountered in practice.

To simulate nonresponse, we used five different nonresponse mechanisms, each defined by independent Bernoulli (Θ_k) trials, where the probability of non-response Θ_k for unit k was specified as follows:

- (M1) Θ_k is constant and independent for all $k \in U$. This is the uniform response mechanism, therefore unconfounded.
- (M2) Θ_k is a decreasing function of x_k specified as $\Theta_k = \exp(-\gamma x_k)$. This is an unconfounded mechanism.
- (M3) Θ_k is an increasing function of x_k specified as $\Theta_k = 1 - \exp(-\gamma x_k)$. This is also an unconfounded mechanism.
- (M4) Θ_k is a decreasing function of y_k specified as $\Theta_k = \exp(-\gamma y_k)$. This is a confounded mechanism.
- (M5) Θ_k is an increasing function of y_k specified as $\Theta_k = 1 - \exp(-\gamma y_k)$. This is also a confounded mechanism.

Note that since we assume x and y to be positively correlated, both (M2) and (M4) are mechanisms such that large units respond more often than small units. The smaller units will be underrepresented in the response set r . Conversely, (M3) and (M5) are mechanisms such that small units respond more often than large units. The larger units will be underrepresented in the response set r .

The first mechanism corresponds to the naive Assumption I discussed in Section 2. (M2) and (M3) correspond to Assumption II while (M4) and (M5) represent fairly simple examples of the confounded mechanisms discussed in connection with Assumption III. For (M2), (M3), (M4) and (M5), the constant γ was determined in such a way that the average nonresponse probability $\bar{\Theta} = (1/N) \sum_U \Theta_k$, is equal to one of the values 10%, 20%, 30% and 40%. Therefore, for each population, there were $5 \times 4 = 20$ different combinations of nonresponse mechanism and nonresponse rate.

For each of the 12 populations, 1,000 samples of size $n = 30$ were drawn. Then for each realized sample, 50 response sets were generated using independent Bernoulli (Θ_k) trials according to one of the 20 combinations of nonresponse mechanism and nonresponse rate. Thus 50,000 response sets were realized for each of the $12 \times 20 = 240$ combinations resulting from cross-classifying the 12 populations with the 20 combinations of nonresponse mechanism and nonresponse rate.

5. RESULTS

We studied the two uncorrected estimators \bar{y}_r (justified under Assumption I) and $\bar{y}_{\text{raimp}} = \bar{x}_s \bar{y}_r / \bar{x}_r$ (justified under Assumption II) and the 8 corrected estimators $\bar{y}_{ci \cdot s}$ and $\bar{y}_{ki \cdot s}$, $i = 1, \dots, 4$ (justified under Assumption III). (We call both \bar{y}_r and \bar{y}_{raimp} uncorrected even though (2.5) shows that we can view \bar{y}_{raimp} as a corrected version of the naive estimator \bar{y}_r . Recall that our principal aim is to correct the bias of \bar{y}_{raimp} when the mechanism is confounded.)

The performance of the 10 estimators is judged by the magnitudes of the relative bias (RB), the relative root mean square error (RRMSE), and the coverage rate (CVR). The RB and the RRMSE of a point estimator \hat{y}_U for \bar{y}_U are defined respectively as,

$$\text{RB}(\bar{y}) = 100 \times \frac{E_p E_q(\hat{y}_U) - \bar{y}_U}{\bar{y}_U},$$

$$\text{RRMSE}(\bar{y}) = 100 \times \frac{\sqrt{E_p E_q(\hat{y}_U - \bar{y}_U)^2}}{\bar{y}_U}.$$

The expectations $E_p E_q(\hat{y}_U)$ and $E_p E_q(\hat{y}_U - \bar{y}_U)^2$ were estimated by Monte Carlo simulation using the 50,000 realized response sets for each of 240 combinations. With this number of replicates, the Monte-Carlo error was less than 0.1%, assuming that the distribution of the \hat{y}_U 's is approximately normal. We will use the abbreviation ARB to denote the absolute relative bias, $|\text{RB}(\bar{y})|$.

We will also discuss the coverage rate (CVR) of the 95% confidence interval constructed as

$$\hat{y}_U \pm 1.96 \sqrt{\hat{V}(\hat{y}_U)}, \quad (5.1)$$

where \hat{y}_U is one of the 10 estimators and $\hat{V}(\hat{y}_U)$ the corresponding variance estimator. For \bar{y}_{raimp} and the 8 corrected estimators, we used the variance estimators described in Section 3. For \bar{y}_r , we used the variance estimator

$$\hat{V}(\bar{y}_r) = \left(\frac{1}{m} - \frac{1}{N} \right) \sum_r (y_k - \bar{y}_r)^2 / (m - 1).$$

The CVR is calculated as 100 times the proportion of the 50,000 response sets such that the interval computed in the manner of (5.1) includes the true mean \bar{y}_U .

For the following discussion, we group the corrected estimators into two groups: s -corrected estimators, which are based on correction factors involving \bar{x}_s or \bar{w}_s , that is, c_2 , c_4 , k_2 and k_4 and r -corrected estimators, which are based on correction factors involving \bar{x}_r or \bar{w}_r , that is, c_1 , c_3 , k_1 and k_3 .

The nonresponse mechanism is the key to the performance of the various estimators. Therefore, Tables 2 and 3 show the behavior of the estimators separately for each of the five mechanisms. We noted that the correlation level and the nonresponse rate do not have a very pronounced effect on the ranking of the estimators. Thus the performance measures ARB, RRMSE and CVR were averaged over 12 cases (three correlation levels \times four nonresponse rates). These averages are shown in Table 2 for the RATIO type regression and in Table 3 for the CONCAVE, CONVEX and NONRATIO regression types.

Table 2

Average ARB, RRMSE (RM) and CVR of Ten Different Estimators for the RATIO Type Populations

For each mechanism, 12 cases were averaged (four nonresponse rates \times three correlation levels)

	M1 (uniform)			M2 (decreasing- x)			M3 (increasing- x)			M4 (decreasing- y)			M5 (increasing- y)		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
\bar{y}_r	0.2	13.9	92.5	12.9	19.1	86.0	9.5	16.5	81.1	19.1	23.6	72.3	14.9	19.9	68.2
\bar{y}_{raimp}	0.2	12.3	92.7	0.6	11.8	93.0	0.4	12.9	92.4	5.3	13.0	92.5	6.0	13.9	85.6
$\bar{y}_{c2 \cdot s}$	1.0	13.3	92.4	4.4	12.6	88.9	8.9	18.3	93.0	1.8	11.8	92.4	3.6	15.3	92.2
$\bar{y}_{c4 \cdot s}$	0.9	13.2	92.3	4.7	12.6	88.6	8.4	17.7	93.0	1.7	11.7	92.3	3.4	14.9	92.2
$\bar{y}_{k2 \cdot s}$	1.1	13.2	92.8	2.4	12.0	90.9	8.0	18.5	93.5	1.7	11.7	93.3	2.2	15.3	92.0
$\bar{y}_{k4 \cdot s}$	1.0	13.1	92.7	2.6	12.0	90.8	7.3	17.7	93.5	1.6	11.7	93.2	1.8	14.7	91.9
$\bar{y}_{c1 \cdot s}$	1.7	14.7	91.4	5.9	13.4	86.4	15.7	26.2	87.6	1.9	12.2	90.9	8.9	21.3	89.8
$\bar{y}_{c3 \cdot s}$	1.6	14.4	91.4	6.2	13.5	86.1	14.9	25.1	87.8	2.1	12.2	90.7	8.3	20.4	90.0
$\bar{y}_{k1 \cdot s}$	2.0	14.7	92.3	3.1	12.3	90.0	15.9	29.6	88.9	1.1	11.7	92.8	8.3	23.8	90.7
$\bar{y}_{k3 \cdot s}$	1.7	14.3	92.3	3.2	12.4	89.8	14.6	27.6	89.3	1.0	11.7	92.7	7.1	21.9	91.0

Table 3

Average ARB, RRMSE (RM) and CVR of Six Different Estimators for CONCAVE, CONVEX, and NONRATIO Populations

(For each mechanism, 12 cases are averaged as in Table 2)

	M1			M2			M3			M4			M5		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
CONCAVE															
\bar{y}_r	0.2	10.4	92.9	10.5	14.8	82.3	7.3	12.7	82.3	12.3	16.0	78.3	8.7	13.4	78.8
\bar{y}_{raimp}	0.2	9.4	94.5	1.4	9.1	93.4	2.6	10.5	94.9	1.9	9.2	94.9	2.1	9.7	92.9
$\bar{y}_{c2 \cdot s}$	1.1	11.4	92.4	6.3	11.4	84.7	11.8	18.8	88.4	3.2	10.2	90.0	5.5	14.2	92.3
$\bar{y}_{c4 \cdot s}$	1.0	11.1	92.8	6.6	11.5	84.3	11.4	18.0	88.8	3.6	10.3	89.8	5.5	13.7	92.7
$\bar{y}_{k2 \cdot s}$	1.0	10.7	93.7	4.5	10.1	89.1	9.5	16.8	91.6	1.7	9.3	93.0	3.7	12.8	93.7
$\bar{y}_{k4 \cdot s}$	0.9	10.5	93.8	4.6	10.1	89.0	9.0	16.0	91.8	1.8	9.3	92.8	3.5	12.3	93.9
CONVEX															
\bar{y}_r	0.9	23.7	90.9	19.0	31.6	92.3	15.0	26.5	76.1	33.2	41.7	76.4	37.1	41.4	37.5
\bar{y}_{raimp}	0.6	21.4	90.6	5.8	21.7	92.8	7.0	22.1	85.6	14.0	25.0	90.0	27.6	33.5	52.0
$\bar{y}_{c2 \cdot s}$	1.2	21.1	91.8	0.4	19.8	91.8	2.0	22.2	92.4	7.3	20.8	93.4	17.8	28.2	71.7
$\bar{y}_{c4 \cdot s}$	1.2	21.3	91.5	0.3	19.9	91.5	1.8	22.3	92.4	6.7	20.6	93.4	18.5	28.5	70.5
$\bar{y}_{k2 \cdot s}$	1.6	21.2	91.9	3.0	21.0	92.0	3.0	22.2	92.6	9.8	22.7	91.7	16.2	27.6	74.0
$\bar{y}_{k4 \cdot s}$	1.4	21.3	91.6	2.9	21.0	91.8	2.6	22.0	92.3	9.5	22.7	91.7	17.6	27.7	72.6
NON-RATIO															
\bar{y}_r	0.1	10.7	92.9	9.7	14.6	86.5	7.3	12.6	81.3	11.9	16.1	80.8	8.8	13.5	77.8
\bar{y}_{raimp}	0.2	9.6	94.5	2.1	9.5	92.4	2.6	10.5	95.3	2.1	9.6	94.4	1.6	9.9	93.3
$\bar{y}_{c2 \cdot s}$	1.1	11.4	92.5	7.0	11.9	83.5	11.9	18.8	89.2	2.6	10.0	90.9	5.3	14.5	92.5
$\bar{y}_{c4 \cdot s}$	1.0	11.3	92.4	7.3	12.1	82.8	11.5	18.1	89.4	2.7	10.1	90.6	4.9	13.8	92.7
$\bar{y}_{k2 \cdot s}$	1.3	11.2	93.4	5.0	10.9	86.9	11.3	19.0	90.7	1.3	9.6	92.8	4.7	14.3	93.5
$\bar{y}_{k4 \cdot s}$	1.1	10.9	93.4	5.2	11.1	86.5	10.6	17.8	91.1	1.3	9.7	92.6	4.1	13.4	93.8

We now comment on the tables. A conclusion of general character is that the respondent mean \bar{y}_r has, as expected, a large bias and a very poor CVR for all of the nonuniform mechanisms. Its performance is satisfactory only for the uniform mechanism (M1). Thus we can focus on the comparisons between the uncorrected \bar{y}_{raimp} on the one hand and the eight corrected estimators on the other. For both of the criteria ARB and RRMSE, we noted that the s -corrected estimators generally gave better results than the r -corrected ones. This is clearly seen in Table 2, where s -corrected and r -corrected estimators are displayed in two separate groups. Given this better behavior of the s -corrected group, we deleted the r -corrected group in Table 3.

5.1 RATIO Type Regression

From Table 2, we draw the following conclusions.

(i) The mechanism (M1) (uniform nonresponse).

When the mechanism (M1) holds, the uncorrected estimator \bar{y}_{raimp} is essentially bias free, and there is no need to correct. However, if the analyst, suspecting a confounded mechanism, has nevertheless chosen one of the corrected estimators, the penalty is not severe. The eight corrected estimators show only a small increase in ARB and in RRMSE compared to \bar{y}_{raimp} .

(ii) The mechanisms (M2) and (M3) (unconfounded, nonuniform and x -value dependent).

For these mechanisms, the ARB is seen to be very small for the uncorrected estimator \bar{y}_{raimp} , as theory would lead us to expect. Our interest is instead focused on the behavior of the eight corrected estimators, since it is important to know if a penalty is associated with an incorrect decision to use one of these estimators. Such a decision would be brought about by an incorrect assumption that the response mechanism is confounded (when in fact it is unconfounded but nonuniform). Table 2 shows that there is indeed some penalty in the form of both increased ARB and increased RRMSE. The penalty is less severe for the s -corrected group. For both groups, the penalty is less severe for the mechanism (M2) than for the mechanism (M3).

(iii) The mechanism (M4) (confounded and y -value dependent).

For this mechanism, a striking feature of Table 2 is that all eight corrected estimators give a substantial bias reduction compared to the uncorrected estimator \bar{y}_{raimp} (and a very large reduction relative to the naive estimator \bar{y}_r). The corrected estimators also show some improvement in RRMSE compared to \bar{y}_{raimp} . The s -corrected estimators perform better than the r -corrected ones. Within the s -corrected group of estimators, the differences are minor, as is the case within the r -corrected group.

(iv) The mechanism (M5) (confounded and y -value dependent).

Table 2 shows that the s -corrected estimators have a smaller ARB than the uncorrected \bar{y}_{raimp} ; their RRMSE is slightly higher. By contrast, the r -corrected estimators "overcorrect" so that both the ARB and the RRMSE exceed the levels observed for \bar{y}_{raimp} . The r -corrected group does not perform well for this mechanism.

In summary, Table 2 shows that if the ratio model (2.2) holds and the assumption of a confounded mechanism is correctly made, the decision to use one of the corrected estimators may lead to a reduced bias. The main difficulty facing the analyst is to accurately predict the nature of the response mechanism causing nonresponse. In particular, it may be difficult for the analyst to separate a confounded mechanism (e.g., one with $\Theta_k = e^{-\gamma y_k}$) from a similar nonuniform unconfounded mechanism (e.g., one with $\Theta_k = e^{-\gamma x_k}$). Yet this subtle difference has a marked effect on the bias of \bar{y}_{raimp} and on the decision whether or not to use a corrected estimator. When the nonuniform unconfounded type applies, we have seen that there is a penalty associated with the corrected estimators, in particular with the r -corrected group.

5.2 Other Regression Types

Table 3 shows the performance of six estimators (the two uncorrected and the four s -corrected) for the CONCAVE, CONVEX, and NONRATIO regression types. As in Table 2, there is little to choose between the estimators when the uniform mechanism (M1) holds. For the two confounded mechanisms, the results in Table 3 do not send a clear message that s -corrected estimation should be attempted even if the assumption of a confounded mechanism is correctly made. Compared to the uncorrected \bar{y}_{raimp} , the s -corrected estimators show a clearly improved performance (in terms of smaller ARB and smaller RRMSE) only for the CONVEX population type. Even in this case, a substantial bias remains after the attempt at correction. For the two unconfounded nonuniform mechanisms (M2) and (M3), it is *a priori* clear that one would not expect improved performance on the part of the s -corrected estimators when compared to \bar{y}_{raimp} . Oddly enough however, we find that the s -corrected estimators work very well for the CONVEX population. These conclusions leave the analyst with a difficult choice if a RATIO type population cannot be assumed. Then it is difficult on the basis of our findings to recommend the use of one of the corrected estimators.

5.3 Coverage Rates

Tables 2 and 3 also show that the variance estimation procedure suggested in Section 3 generally works well. Indeed the coverage rates for the corrected estimators are uniformly good whenever the ARB is small. In particular,

Table 4

Average ARB, RRMSE (RM) and CVR of the Two Uncorrected Estimators and the c_4 - and k_4 - Corrected Estimators
(Averaged Over All Population Types)

	M1			M2			M3			M4			M5			Overall		
	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR	Av. ARB	Av. RM	Av. CVR
\bar{y}_r	0.3	14.7	92.3	13.0	20.0	86.8	9.8	17.1	80.2	19.1	24.4	77.0	17.4	22.1	65.6	11.9	19.6	80.4
\bar{y}_{raimp}	0.3	13.2	93.1	2.5	13.0	92.9	3.1	14.0	92.0	5.8	14.2	93.0	9.3	16.7	81.0	4.2	14.2	90.4
$\bar{y}_{c4 \cdot s}$	1.0	14.2	92.3	4.7	14.0	86.8	8.3	19.0	90.8	3.7	13.2	91.5	8.1	17.7	87.0	5.2	15.6	89.7
$\bar{y}_{k4 \cdot s}$	1.1	14.0	92.9	3.8	13.6	89.5	7.4	18.4	92.2	3.6	13.3	92.6	6.7	17.0	88.0	4.5	15.2	91.0

for the unconfounded mechanisms (M2) and (M3), the coverage rates for the corrected estimators are about equal to or better than those for the uncorrected estimators.

5.4 Overall Comments

From the summary Table 4, we note that, as expected, \bar{y}_r and \bar{y}_{raimp} show the best performance for the uniform response mechanism (M1). The uncorrected estimator \bar{y}_{raimp} is the best one for the unconfounded mechanisms (M2) and (M3), while the corrected estimators are the best ones for the confounded mechanism (M4) and (M5).

Finally, on the average over all 240 cases included in our study, we note from the overall column of Table 4 that \bar{y}_{raimp} and $\bar{y}_{k4 \cdot s}$ perform similarly with the former having a slightly smaller bias and the latter having slightly better coverage rate.

6. CONCLUSIONS

It has long been recognized that nonresponse causes bias in survey estimates, except in rare cases. Imputation is a widely used practice to handle nonresponse, because it is convenient to work with a complete data set. There are many imputation rules as well as some softwares that can be used in large scale surveys. Imputation is sometimes applied without critical questioning, and, although widely used, imputation does not solve the critical problem of bias caused by nonresponse.

In this paper, we have examined ratio imputation. The ordinary ratio imputation $\hat{B}_{r \cdot x_k}$ is justified (that is, it produces no bias) if two conditions hold: (a) the regression model behind the ratio imputation rule holds (that is, a linear regression through the origin); (b) the response mechanism is unconfounded.

The results of our simulation give some idea of the magnitude of the bias of the usual ratio imputation estimator \bar{y}_{raimp} when one or both of the two conditions break down. We considered several nonuniform response mechanisms, confounded as well as unconfounded mechanisms. We also considered breakdown of the regression model behind ratio imputation.

We argued that a confounded mechanism can sometimes be realistically assumed in a survey. We showed that if an assumption of confounded response mechanism is correctly made, and if the model behind the ratio imputation is valid, one can make some progress toward bias reduction using the s -corrected estimators in this paper. They have substantially less bias than the uncorrected estimator \bar{y}_{raimp} . The s -corrected estimators are generally more effective than the r -corrected estimators for reducing the bias.

Suppose the analyst is working under the assumption that the ratio model (2.2) holds. Our simulation study then leads to suggested estimators according to the following Table 5, depending on the assumed nature of the response mechanism and on the nonresponse rate. The entry "any" means any of the 10 estimators in Table 2.

Table 5

Suggested Estimators for Each Nonresponse Mechanism

Nonresponse Rate	Suggested Estimator		
	Response Mechanism		
	Uniform	Unconfounded	Confounded
($\leq 10\%$)	any	any but \bar{y}_r	any but \bar{y}_r
(> 10%)	any ¹	\bar{y}_{raimp}	s -corrected

Note 1: \bar{y}_{raimp} as a slight advantage over the others.

If the regression model behind ratio imputation fails, the situation is less clear. Unless the naive assumption of a uniform response mechanism holds (which is unlikely), the uncorrected ratio imputation estimator \bar{y}_{raimp} can have considerable bias. We found that \bar{y}_{raimp} is particularly prone to bias for the CONVEX type population where the s -corrected group of estimators usually have smaller bias than \bar{y}_{raimp} . On the other hand, for the CONCAVE and the NONRATIO type populations, \bar{y}_{raimp} is generally more resistant to bias than the s -corrected estimators.

7. ACKNOWLEDGMENT

The authors wish to thank the referees and the associate editor for their helpful comments. An earlier version of this paper was presented at the Annual Research Conference (ARC) in Arlington, Virginia, March 22-25, 1992.

REFERENCES

- BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.
- FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.
- FAY, R.E. (1989). Estimating nonignorable nonresponse in longitudinal surveys through causal modeling. In *Panel Surveys* (Eds. D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh), 375-399.
- GREENLESS, J.S., REECE, W.S., and ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 82, 251-261.
- LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- RUBIN, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.
- RUBIN, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12, 37-47.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- SÄRNDAL, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, 337-347.
- SÄRNDAL, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Dual System Estimation of Census Undercount in the Presence of Matching Error

YE DING and STEPHEN E. FIENBERG¹

ABSTRACT

Dual system estimation (DSE) has been used since 1950 by the U.S. Bureau of Census for coverage evaluation of the decennial census. In the DSE approach, data from a sample is combined with data from the census to estimate census undercount and overcount. DSE relies upon the assumption that individuals in both the census and the sample can be matched perfectly. The unavoidable mismatches and erroneous nonmatches reduce the accuracy of the DSE. This paper reconsiders the DSE approach by relaxing the perfect matching assumption and proposes models to describe two types of matching errors, false matches of nonmatching cases and false nonmatches of matching cases. Methods for estimating population total and census undercount are presented and illustrated using data from 1986 Los Angeles test census and 1990 Decennial Census.

KEY WORDS: Capture-recapture; Matching bias; Modelling matching error; Multinomial likelihood.

1. INTRODUCTION

The problem of undercount in the U.S. census has been of special concern since the first census of 1790 (Jefferson 1986). The DSE (or capture-recapture) approach has been used in conjunction with the census to evaluate population coverage as part of what is called the post-enumeration survey (PES) program. Ericksen and Kadane (1985) and Wolter (1986) describe the use of the DSE approach in the context of the 1980 decennial census. A new design for the PES was planned for the 1990 decennial census and refinements in methodology were examined in connection with a 1986 test census in central Los Angeles County, referred to as the Test of Adjustment Related Operations (TARO). Diffendal (1988) discusses methodology, operations, and the results of TARO, and Hogan and Wolter (1988) and Schenker (1988) provide evaluation of the operations and assumptions underlying the DSE approach.

The PES approach to dual-system estimation uses two samples, called the P-sample and the E-sample. The P-sample which is drawn separately from the census, helps to measure census omissions; the E-sample drawn from the census enumerations, helps to measure census erroneous enumerations. For the 1986 TARO, the dual-system estimator for the population size, N , which combines the information from the P-sample and the E-sample takes the form:

$$\hat{N} = (\text{CEN} - \text{EE} - \text{SUB}) \cdot N_p / M,$$

where CEN is the unadjusted census count; EE is the estimated number of erroneous enumerations and unmatchable

persons included in the census; SUB is the number of whole-person substitutions in the census; N_p is the number of people in the P-sample; M is the estimate of the number of people in both census and the P-sample. For details see Diffendal (1988) or Wolter (1986). For the variation on this formula as used in conjunction with the 1990 census, see Hogan (1992, 1993).

DSE and the matching problem gained considerable attention in the 1970's due to its use in estimating births and deaths in developing countries, and it is thought by some that perhaps the greatest problem with the dual-system estimation approach used in 1980 census was the rate of matching error (Fienberg 1989). Jaro (1989) describes the technological innovations for matching introduced by the Bureau of the Census for 1990 and the test of the related matching methodology in a 1985 pre-test. Biemer (1988) considers models for evaluating the impact of matching error on estimates of census coverage error without attempting to correct for the matching bias in the usual dual-system estimate. The actual procedure used in the 1990 census included not only a computer matching algorithm and various clerical follow-ups but also logistic regression models for unresolved cases in both the P-sample and E-sample (see Belin *et al.* 1993).

Matching is used to determine the census enumeration status of the people enumerated in the P-sample. Specifically, those people in the P-sample who are matched to the census are considered to have been enumerated. People in the P-sample who do not match are, for the most part, considered to have been missed by the census. Matching errors can occur for two general reasons:

¹ Ye Ding is Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg is Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

1. The information reported by the respondents/interviewers was incorrect.
2. Correct information was reported, but it was not correctly used.

Moreover, two types of errors can occur: false matches of nonmatching cases and false nonmatches of matching cases. False matches of nonmatching cases may be divided into

- (a) instances in which a P-sample case was erroneously matched to the enumeration of another person, but a match to that actual E-sample case should have been made, and
- (b) instances in which no match should have been made.

The former case is not “serious” for the purposes of estimating N , since such false matches would have been, in fact, correctly classified as a match to the census. In the second case, however, the number of nonmatches becomes understated. False nonmatches to the census, on the other hand, have the effect of overestimating the nonmatch rate. Fay, Passel, Robinson and Cowan (1988) note that false nonmatches probably represent a greater concern than false matches. False matches are less common than false nonmatches because matches can be reviewed easily.

In Section 2, we propose models for matching errors and then, in Section 3 and 4, we present a systematic procedure for the estimation of the population total and thus the census undercount. In Section 5, we analyze the data from 1986 Los Angeles test census and 1990 Decennial Census to show how our method accounts for matching errors in the undercount estimates.

2. MODELING MATCHING ERRORS

For simplicity, we assume that the matching mechanism is constrained, in the sense that no individual in one sample can be matched with more than one individual in another sample. Moreover, we implicitly assume a version of simple random sampling, within strata, and this yields a standard multinomial sampling model for dual system estimation. This simplification allows us to focus on the impact of matching and its mechanisms. In what follows, we provide a way to view the recapture data, for the purpose of setting up models for matching.

Let $Z_{N \times 1}$ be the characteristic vector for the whole population, such that the i -th component of $Z_{N \times 1}$ contains the characteristics for the i -th individual, where $1 \leq i \leq N$. Not all the components in $Z_{N \times 1}$ can be observed in any one sample. The object is to estimate N , the size of the population, from information from two samples. One could view drawing a sample from the population as drawing some components in $Z_{N \times 1}$ at random to form a new vector Y . Then, missing or misreporting of certain characteristics in those components drawn may cause matching errors. Henceforth we will refer to the first

sample as Y_1 and the second sample as Y_2 , and in the following discussion they will be the two capture-recapture samples for dual system estimation.

Two types of matching errors can occur: false non-matches of matching cases, and false matches of non-matching cases. We will refer to the former as a type 1 error and the latter as a type 2 error. We can focus on modeling one or both types of error. Under perfect matching, each component in Y_1 or Y_2 contains the same information as in $Z_{N \times 1}$, and the number of matches will be the number of elements common to Y_1 and Y_2 . When faced with uncertain matching, we consider the following simple model:

Model (A):

- (i) Assume that those matched pairs of components under perfect matching will still be matched, each with common probability α , $0 < \alpha \leq 1$.
- (ii) All those unmatched will remain unmatched, *i.e.*, no false matches.

Model (A) characterizes a mechanism for type 1 matching error with error probability $1 - \alpha$, assuming that type 2 matching error is negligible.

To develop a model for both types of matching error, we need to consider carefully all the possibilities that lead to false matches. When there is no matching error, one can write $Y_1 = (M_1, N_1)$ and $Y_2 = (M_2, N_2)$, so that sets M_1 and M_2 have the same size and every individual in M_1 is correctly matched with one individual in M_2 and vice versa, N_1 is the set of those in sample Y_1 who are not matched with any one in sample Y_2 , and N_2 is the set of those in sample Y_2 who are not matched with any one in sample Y_1 . When matching errors are present, false matches can occur in the following ways:

- (a) A person in M_1 is matched incorrectly with a person in M_2 .
- (b) A false match occurs between M_1 and N_2 .
- (c) A false match occurs between M_2 and N_1 .
- (d) A false match occurs between N_1 and N_2 .

We note that each of (a), (b), (c) happens only when at least 2 errors are made, that is, the correct match is not made and an incorrect match is made. Since such errors occur with small probability, we assume for simplicity that cases (a), (b), (c) have negligible probability of occurrence in the next model.

Model (B):

- (i) Assume, as in model (A), that matching pairs between M_1 and M_2 will still be matched, but with probability α , $0 < \alpha \leq 1$.
- (ii) Assume that false matches of types (a), (b), (c) are negligible.
- (iii) Assume that each person in N_1 will be matched with someone in N_2 with a common probability β , $0 \leq \beta < 1$.

Even though, in theory, both α and β can vary from 0 to 1, in the census context we expect that $\alpha \approx 1$, and $\beta \approx 0$.

We can also consider instances in which the matching error probabilities and capture probabilities potentially vary over identifiable population subgroups. In other words, the population can be divided into strata, by demographic (e.g., age, race, sex) and geographic variables, within which the matching error probabilities and capture probabilities could be assumed to be more homogeneous than in the whole population. Suppose the whole population consists of l strata. Let $Z_{N_i \times 1}^i$ be the characteristic vector for the population of the i -th stratum with unknown size N_i , and let Y_{i1} , Y_{i2} be two samples taken from the i -th stratum which are used to get an estimate \hat{N}_i . Then we can form an estimate of the overall population size by setting $\hat{N} = \sum_{i=1}^l \hat{N}_i$. We can refine models (A) and (B) as follows:

Model (A'):

Assume model (A) holds within each stratum, and let α_i be the probability of a match for matching components in stratum i , $0 < \alpha_i \leq 1$, $1 \leq i \leq l$.

Model (B'):

Assume model (B) holds within each stratum, and let the two probability parameters for i -th stratum be α_i , β_i , $1 \leq i \leq l$.

For 1990 PES, the P-sample matching was conducted using the sample blocks plus a ring of surrounding blocks (Hogan 1993). Geocoding errors may lead to false matches across geographically defined post-strata, and false matches are possible for demographically defined post-strata. Models (B') implicitly assumes that there are no false matches across post-strata. Further, all of the models represent a simplification of the underlying sample design of the PES.

3. ESTIMATE THE POPULATION TOTAL

In this section, we consider estimation of the population total under the various matching models, (A), (A'), (B), and (B'), assuming the validity of usual assumptions of independence of the two samples and homogeneous probabilities of inclusion in the samples. For models involving heterogeneous catchability and/or dependence, see the three-sample approach in Darroch *et al.* (1993) and the approach in Alho *et al.* (1993).

Let N be the number of individuals in the population under consideration, x_{1+} the number of individuals in Y_1 , x_{+1} the number of individuals in Y_2 , and x_{11} the number of individuals in both samples. The number of individuals observed in Y_2 but not Y_1 is $x_{21} = x_{+1} - x_{11}$ and the number observed in Y_1 but not Y_2 is $x_{12} = x_{1+} - x_{11}$.

One can arrange the capture-recapture data in a 2×2 contingency table with one missing cell:

		Sample Y_2	
		present	absent
Sample Y_1	present	x_{11}	x_{12}
	absent	x_{21}	—

where we use symbol “—” to indicate the missing cell, and standard notation for marginal totals: $x_{1+} = x_{11} + x_{12}$, $x_{+1} = x_{11} + x_{21}$. There is a corresponding 2×2 table of probabilities, $p_{ij} = \Pr[\text{any individual falls into } (i,j) \text{ cell}]$,

		Sample Y_2	
		present	absent
Sample Y_1	present	p_{11}	p_{12}
	absent	p_{21}	p_{22}

with the usual linear constraint

$$\sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1.$$

Let n be the number of observed different individuals in the two samples, i.e., $n = x_{11} + x_{12} + x_{21}$. If we assume that the samples are randomly selected with homogeneous selection probabilities, then the numbers of individuals in the four cells have a multinomial distribution

$$(x_{11}, x_{12}, x_{21}, N - n) \sim \text{Mult}(N, p_{11}, p_{12}, p_{21}, p_{22}).$$

We use the conditional likelihood approach developed by Sanathanan (1972). For fixed n , (x_{11}, x_{12}, x_{21}) has a multinomial distribution with likelihood function

$$L_1(p_{11}, p_{12}, p_{21}) = \frac{n!}{x_{11}! x_{12}! x_{21}!} \cdot \frac{p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}}}{(p_{11} + p_{12} + p_{21})^n}. \quad (1)$$

Then n is viewed as being binomially distributed with sample size N and probability $p_{11} + p_{12} + p_{21}$, and the corresponding likelihood is

$$L_2(N) = \frac{N!}{n! (N - n)!} (p_{11} + p_{12} + p_{21})^n [1 - (p_{11} + p_{12} + p_{21})]^{N-n}. \quad (2)$$

In the conditional approach we derive maximum likelihood estimates for the cell probabilities based on the likelihood (1), then find the value of N which maximizes (2), given

the values of the cell probabilities. Sanathanan (1972) has shown that under suitable regularity conditions both conditional and unconditional likelihood estimates of N are consistent and have the same asymptotic multivariate normal distribution. The conditional approach is particularly suitable for a large sample problem like ours.

Under the equal catchability assumption, we let p_1 be the probability that any individual in the population is included in Y_1 , and similarly we let p_2 be the probability of inclusion in Y_2 . The probabilities p_1 and p_2 are usually referred to as capture probabilities and they do not depend on how the matching mechanism operates. Then the probability that an individual is in both samples is $p_1 p_2$, and the probability of being in set N_1 is $p_1(1 - p_2)$. Since model (A) is a special case of model (B) with $\beta = 0$, we focus on formulating the problem under model (B). To do this, we first need to work out the parametric specification of the cell probabilities. An individual will fall into the (1, 1) cell in the 2×2 table only in two cases, *i.e.*, the individual is actually in both samples and a match is made, or, using the notation in the last section, an individual who is actually in N_1 is incorrectly matched with some one in N_2 . Here the matching direction from N_1 to N_2 is implicitly assumed in (iii) of model (B). The probability that the former case occurs is $\alpha p_1 p_2$, and the probability that the latter case occurs is $\beta p_1(1 - p_2)$. Furthermore, the two cases are mutually exclusive. Thus, we have $p_{11} = \alpha p_1 p_2 + \beta p_1(1 - p_2)$, and, $p_{12} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1(1 - p_2)$, $p_{21} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1(1 - p_2)$. Rao (1957) studied regularity conditions under which there exist unique maximum likelihood estimates of parameters in a multinomial distribution. His conditions are satisfied by the parameterization of $\{p_{ij}\}$ here.

For $\alpha = 1, \beta = 0$, this setup reduces to the usual two sample problem and there exist well known solutions in closed form for resulting likelihood equations for the conditional likelihood (1) (*cf.* Bishop *et al.* 1975, chap. 6, p. 232), leading to the usual dual-system estimator, $\hat{N}_{DSE} = x_{1+}x_{+1}/x_{11}$. Otherwise, the maximum likelihood estimates cannot be written in closed form. Once we have \hat{p}_1 and \hat{p}_2 , however, the conditional maximum likelihood estimates for p_1 and p_2 , the conditional maximum likelihood estimate for N can be written as

$$\hat{N} = \frac{n}{\hat{p}_1 + \hat{p}_2 - (\alpha - \beta)\hat{p}_1\hat{p}_2 - \beta\hat{p}_1}, \quad (3)$$

(*cf.* Chapman 1951). Under model (A') or (B'), for the i -th stratum, one can use the estimates of the parameters computed under model (A) or (B) for the data of that stratum, and then sum over strata for an estimate of the population total.

4. ESTIMATE MATCHING ERROR RATES BY REMATCH STUDY DATA

In what follows, we give estimates of the matching error rate parameters α and β using the data from the Matching Error Study (rematch study), one of the operations conducted by the Census Bureau in the 1986 Los Angeles test census to evaluate the PES. Briefly, the rematch typically operates for a sample of cases, using more extensive procedures, highly qualified personnel and reinterviews to obtain estimates of the bias associated with the previous matching process. For further details, see Childers, Diffendal, Hogan and Mulry (1989). In their discussion of the Matching Error Study in Los Angeles TARO, Hogan and Wolter (1988) state that "The rematch was done independently of the original match, and the discrepancies between the match and the rematch results are adjudicated. Because of this intensive approach to the rematch, we believe the rematch results represent true match status, while differences between the match and rematch results represent the bias in the original match results."

The data collected in a rematch study can be displayed as in the following table

		Rematch Study Data	
		Rematch Classification	
		Matched	Not Matched
Original Classification	Matched	y_{11}	y_{12}
	Not Matched	y_{21}	y_{22}

To estimate α and β , we assume that in the original matching process, errors are made according to model (B) and that errors in the rematch process can be disregarded, *i.e.*, the rematch is assumed to be perfect. It then follows that $y_{11} + y_{21}$ is the true number of matches, and thus is fixed, while y_{11} is a random variable having a binomial distribution, *i.e.*, $y_{11} \sim \mathcal{B}(y_{11} + y_{21}, \alpha)$. Thus the maximum likelihood estimate of α is $\hat{\alpha} = y_{11}/(y_{11} + y_{21})$, and the maximum likelihood estimate of the false nonmatch rate γ is $\hat{\gamma} = 1 - \hat{\alpha} = y_{21}/(y_{11} + y_{21})$. By the same argument, $y_{12} \sim \mathcal{B}(y_{12} + y_{22}, \beta)$, and the maximum likelihood estimate of the false match rate is $\hat{\beta} = y_{12}/(y_{12} + y_{22})$.

We can use the estimates of the matching error rates derived here to analyze the data from the rematch study from the Los Angeles test census. Very often, in addition to estimating the size of a population, it is of interest to estimate the size of a subpopulation such as black, white, or a subpopulation at a certain geographical location. In such case, it is more appropriate to allow for heterogeneity

of matching error rates across various population strata by using estimates of matching error rates for each stratum of interest. Such estimates can be obtained by conducting a rematch study within each stratum and then using the derived estimates. Data for applying model (B') are available from 1990 Census and are analyzed here.

5. APPLICATIONS

5.1 Application of One Stratum Model to 1986 TARO

Hogan and Wolter (1988) present the rematch data from the 1986 Los Angeles TARO. The rematch results for the P-sample are given in Table 1 in the form of a cross-tabulation of match statuses as assigned from the original TARO match and the rematch. Table 2 presents the two way table of data for the 1986 TARO, with no post-stratification. The estimate of the number missed by both systems, 5,870 is approximately the same order of magnitude as census substitutions 5,259 and erroneous enumerations 6,426 (Hogan and Wolter 1988). Rematch results for the E-sample are presented in Table 3. Let CP, EP be the total correct enumeration and erroneous enumeration by production classification, and let CR, ER be the total correct enumeration and erroneous enumeration by rematch classification, then based on the data in Table 3, Hogan and Wolter (1988) conclude that the original rate of erroneous enumerations (EE), $EP/(CP + EP) = 325/(325 + 19,269) = .016$ should be increased to about $ER/(CR + ER) = 411/(411 + 19,334) = .021$.

Table 1

Results of 1986 Los Angeles Test Census Rematch Study:
P-Sample. Source: Hogan and Wolter (1988)

Original Match Classification	Rematch Classification			Total
	Matched	Not Matched	Un-resolved	
Matched	16,623	18	55	16,696
Not matched	88	2,164	56	2,308
Unresolved	17	0	132	149
Total	16,728	2,182	243	19,153

Table 2

Data and Dual-System Estimate for 1986 Los Angeles Test Census. Source: Hogan and Wolter (1988)

		PES		
		Counted	Missed	Total
Correct Census Enumerations*	Counted	298,204	45,463	343,667
	Missed	38,503	5,870	44,373
	Total	336,707	51,333	388,040

* Correct Enumerations = Total Census Enumerations - Substitutions - Erroneous Enumerations.

Table 3

Results of 1986 Los Angeles Test Census Rematch Study:
E-Sample. Source: Hogan and Wolter (1988)

Original Match Classification	Rematch Classification			Total
	Correct Enumeration	Erroneous Enumeration	Unresolved	
Correct enumeration	19,153	28	88	19,269
Erroneous enumeration	41	283	1	325
Unresolved	140	100	223	463
Total	19,334	411	312	20,057

We now reanalyze the data in Table 2 using model (B), but ignoring the unresolved cases in Table 1 because their classification status are unavailable to us. From the data in Table 1 we estimate $\hat{\gamma} = 1 - \hat{\alpha} = 88/(16,623 + 88) = .53\%$, and $\hat{\beta} = 18/(18 + 2,164) = .82\%$. In Table 4, we present the estimates and associated standard deviations under model (B) and those from the traditional DSE. The standard deviations are computed using asymptotic normality, for details, see Ding (1990, 1993a, 1993b). The estimated undercount is then defined to be undercount = $(\hat{N} - \text{CEN})/\hat{N} \times 100\%$, and CEN is the total census enumerations, *i.e.*, $\text{CEN} = \text{Correct Census Enumeration} + \text{Substitutions} + \text{EE} = 343,667 + 5,259 + 6,426 = 355,352$. The estimates on the last row of Table 4 indicates that the undercount estimate provided by the DSE should be reduced by $8.42\% - 8.05\% = .37\%$. We recall that Hogan and Wolter (1988) argue that the original rate of EE should be increased by $2.1\% - 1.6\% = .5\%$ as a result of information in the rematch study. This then gives an additional adjustment to the estimated undercount of about .5%. Overall, we estimate that the undercount estimate was biased upward by about .9% (assuming the overlapping is negligible, even though two components are not strictly additive).

Table 4

Comparison of Estimates for 1986 Los Angeles Test Census

Parameter	DSE (SD)	MLE from Model (B) (SD)
p_1	.8856 (5.48×10^{-4})	.8892 (5.51×10^{-4})
p_2	.8677 (5.78×10^{-4})	.8712 (5.86×10^{-4})
N	388,040 (87)	386,470 (79)
Undercount (%)	8.42%	8.05%

Table 5

13 Evaluation Post-strata (EPS) for 1990 PES

1	Northeast, Central City, Minority
2	Northeast, Central City, Nonminority
3	U.S., Noncentral City, Minority
4	Northeast, Noncentral City, Nonminority
5	South, Central City, Minority
6	South, Central City, Nonminority
7	South, Noncentral City, Nonminority
8	Midwest, Central City, Minority
9	Midwest, Central City, Nonminority
10	Midwest, Noncentral City, Nonminority
11	West, Central City, Minority
12	West, Central City, Nonminority
13	West, Noncentral City, Nonminority + Indian

Table 6

Dual System Data for 13 EPS of 1990 PES

EPS	x_{1+} (Census)	x_{+1} (P-sample)	x_{11}
1*	5,966,529	4,656,305.09	4,284,132.78
2	9,235,705	8,685,235.79	8,626,362.34
3*	24,255,611	22,628,349.88	21,068,045.55
4	31,173,378	30,150,266.34	29,966,142.62
5*	9,985,055	8,809,620.02	8,249,407.92
6	13,977,529	13,582,482.34	13,278,614.01
7	47,548,548	44,059,397.93	42,987,517.59
8*	4,060,286	3,714,168.27	3,520,314.04
9	11,826,352	10,058,288.52	9,854,052.95
10	39,343,787	38,358,735.32	38,031,852.01
11*	7,283,885	5,743,998.39	5,365,961.67
12	11,073,872	10,512,339.59	10,222,147.69
13	26,415,232	26,721,116.28	26,025,370.25

*Corresponds to minority post-stratum.

Table 7

Results of Rematch Study for 13 EPS of 1990 PES: P-Sample

EPS	y_{11}	y_{21}	y_{12}	y_{22}
1*	14,301	124	31	2,773
2	15,051	36	16	1,136
3*	28,784	293	49	4,166
4	32,753	703	27	2,058
5*	28,674	189	18	3,738
6	21,757	69	36	1,156
7	48,061	47	20	3,278
8*	14,800	58	21	2,527
9	16,527	39	20	874
10	43,721	120	107	1,664
11*	12,522	133	11	2,097
12	15,122	59	8	1,078
13	43,356	232	108	4,583

Table 8

Results of Rematch Study for 13 EPS of 1990 PES: E-Sample

EPS	CP	EP	CR	ER
1*	17,027	1,415	17,106	1,645
2	15,821	879	15,631	932
3*	32,420	2,430	32,322	2,446
4	33,369	1,242	32,922	1,665
5*	32,412	1,880	33,030	2,044
6	24,392	1,225	24,336	1,284
7	51,107	2,908	50,929	3,047
8*	17,174	1,518	17,133	1,526
9	18,279	648	18,228	656
10	44,450	1,604	44,584	1,631
11*	13,644	985	13,693	909
12	15,647	522	15,590	583
13	49,647	2,062	49,545	2,334

5.2 Application of Multiple Strata Model to 1990 Census

We now analyze stratified data from the evaluation of the PES carried out as part of 1990 decennial census. Hogan (1993) describes operations and results for the 1990 PES, Mulry and Spencer (1991, 1993) present total error analysis, and Davis *et al.* (1991) report on the PES Matching Error Study (MES). The MES was conducted for each of 13 Evaluation Post-strata (EPS) by geographic region and ethnic group. Of the 13 EPS listed in Table 5, five correspond to substantial minority populations (Blacks and Hispanics), *i.e.*, EPS 1, 3, 5, 8 and 11. In Table 6, we present the dual system data for each of the 13 EPS, and we give, in Table 7 and Table 8, relevant rematch data for the P-sample and E-sample. These data are drawn from the final reports on PES evaluation projects P7 and P10 by the Census Bureau (Davis and Biemer 1991a, 1991b). The P-sample for the 1990 PES consisted of about 172,000 housing units (Hogan 1992). The P-sample data are weighted to get estimates of x_{+1} (P-sample total) and x_{11} (total matches) in the usual analysis of the dual system data and the analysis presented here. Nevertheless, the actual unweighted P-sample data can be used to make inference, see Appendix for comparison between estimates from actual P-sample data and estimates from weighted P-sample data.

In Table 9, we give the usual dual system estimates and standard deviations of the capture probabilities (*i.e.*, coverage rate by Census or P-sample) for each of the 13 EPS. Estimates in Table 10 indicate that there is significant variation in matching error rates across the EPS. Among three EPS with $\hat{\gamma}$ larger than .01%, EPS 3 and EPS 11 are minority post-strata. This suggests that the nonmatch rate may be higher for minority post-strata than for the remainder. On the other hand, there is no clear evidence from the estimates of $\hat{\beta}$ that the false match rate is higher

Table 9

Usual Dual System Estimates and Standard Deviations
for 13 EPS of 1990 PES

EPS	\hat{p}_1 (SD)	\hat{p}_2 (SD)	\hat{N} (SD)
1*	0.92007 (12.57×10^{-5})	0.71803 (18.42×10^{-5})	6,484,855 (470)
2	0.99322 (2.78×10^{-5})	0.93402 (8.17×10^{-5})	9,298,737 (67)
3*	0.93105 (5.33×10^{-5})	0.86858 (6.86×10^{-5})	26,051,987 (540)
4	0.99389 (1.42×10^{-5})	0.96127 (3.46×10^{-5})	31,364,919 (88)
5*	0.93641 (8.22×10^{-5})	0.82618 (11.99×10^{-5})	10,663,134 (390)
6	0.97763 (4.01×10^{-5})	0.95000 (5.83×10^{-5})	14,297,391 (131)
7	0.97567 (2.32×10^{-5})	0.90408 (4.27×10^{-5})	48,734,156 (359)
8*	0.94781 (11.54×10^{-5})	0.86701 (16.85×10^{-5})	4,283,875 (190)
9	0.97969 (4.45×10^{-5})	0.83322 (10.84×10^{-5})	12,071,466 (224)
10	0.99148 (1.48×10^{-5})	0.96665 (2.86×10^{-5})	39,681,946 (108)
11*	0.93419 (10.35×10^{-5})	0.73669 (16.32×10^{-5})	7,797,041 (443)
12	0.97240 (5.05×10^{-5})	0.92309 (8.01×10^{-5})	11,388,243 (164)
13	0.97396 (3.08×10^{-5})	0.98524 (2.35×10^{-5})	27,121,400 (104)

Table 10

Estimates of Matching Error Rates
for 13 EPS of 1990 PES

EPS	$\hat{\gamma}$ (%)	$\hat{\beta}$ (%)
1*	0.009	0.011
2	0.002	0.014
3*	0.010	0.012
4	0.021	0.013
5*	0.007	0.005
6	0.003	0.030
7	0.001	0.006
8*	0.004	0.008
9	0.002	0.022
10	0.003	0.060
11*	0.011	0.005
12	0.004	0.007
13	0.005	0.023

Table 11

MLEs from Model (B') and Standard Deviations
for 13 EPS of 1990 PES

EPS	\hat{p}_1 (SD)	\hat{p}_2 (SD)	\hat{N} (SD)
1*	0.92406 (12.68×10^{-5})	0.72114 (18.79×10^{-5})	6,456,833 (446)
2	0.99464 (2.79×10^{-5})	0.93536 (8.30×10^{-5})	9,285,474 (92)
3*	0.93896 (5.38×10^{-5})	0.87597 (7.01×10^{-5})	25,832,352 (279)
4	0.99999 (2.65×10^{-5})	0.98070 (3.64×10^{-5})	30,731,889 (781)
5*	0.94166 (8.28×10^{-5})	0.83080 (12.13×10^{-5})	10,603,717 (306)
6	0.97922 (4.03×10^{-5})	0.95154 (6.03×10^{-5})	14,274,182 (64)
7	0.97600 (2.32×10^{-5})	0.90438 (4.30×10^{-5})	48,717,792 (338)
8*	0.95034 (11.59×10^{-5})	0.86933 (17.06×10^{-5})	4,272,459 (159)
9	0.97756 (4.47×10^{-5})	0.83141 (11.12×10^{-5})	12,097,806 (285)
10	0.99217 (1.50×10^{-5})	0.96733 (3.06×10^{-5})	39,654,306 (90)
11*	0.94239 (10.46×10^{-5})	0.74316 (16.58×10^{-5})	7,729,158 (359)
12	0.97561 (5.07×10^{-5})	0.92614 (8.10×10^{-5})	11,350,674 (101)
13	0.97895 (3.10×10^{-5})	0.99029 (2.42×10^{-5})	26,983,168 (355)

Table 12

Undercount Percentage and Bias Estimates
for 13 EPS of 1990 PES

EPS	UC(DSE)	UC(P)	UC(E)	UC(T)	Bias(P)	Bias(E)	Bias(T)
1*	6.40	5.99	5.30	4.89	0.41	1.10	1.51
2	-0.69	-0.83	-1.05	-1.20	0.14	0.36	0.51
3*	5.59	4.79	5.53	4.72	0.80	0.06	0.87
4	-0.11	-2.17	-1.33	-3.39	2.06	1.23	3.29
5*	5.03	4.49	4.68	4.15	0.53	0.35	0.88
6	1.22	1.06	0.99	0.83	0.16	0.23	0.39
7	1.77	1.73	1.50	1.47	0.03	0.26	0.29
8*	3.52	3.26	3.46	3.20	0.26	0.06	0.32
9	1.05	1.26	1.00	1.21	-0.22	0.05	-0.17
10	0.41	0.34	0.36	0.29	0.07	0.05	0.12
11*	5.26	4.43	5.77	4.94	0.83	-0.51	0.32
12	1.89	1.56	1.51	1.19	0.32	0.38	0.70
13	1.79	1.29	1.28	0.78	0.50	0.51	1.01

for minority post-strata, or the other way around. In Table 11, we give maximum likelihood estimates and standard deviations under model (B'). Heterogeneity in the capture probabilities is significant. This heterogeneity together with the variation in the matching error rates suggests that model (B') is more appropriate than model (B). The asymptotic standard deviations in Table 9 and 11 appear unusually small comparing to the sample size of N . Ding (1993b) shows that this is a typical feature of the dual system problem when the capture probabilities are very high, as it is the case in census application. Despite very narrow confidence intervals, simulation studies in Ding (1993b) show that the asymptotic normal approximation being used is highly accurate in terms of coverage probability.

Table 12 provides estimates of matching bias of various sources in the undercount estimate by the usual DSE. UC(DSE) is the undercount estimate from the DSE defined in the same way as for the 1986 TARO estimate; UC(P) is the undercount estimate computed by MLE from matching error model to adjust for matching bias in P-sample, and $\text{Bias(P)} = \text{UC(DSE)} - \text{UC(P)}$. Again, following Hogan and Wolter (1988), we define the bias in E-sample operation by $\text{Bias(E)} = \text{ER}/(\text{CR} + \text{ER}) - \text{EP}/(\text{CP} + \text{EP})$, and the undercount estimate correcting for E-sample error by $\text{UC(E)} = \text{UC(DSE)} - \text{Bias(E)}$. Finally the total matching bias by both P-sample and E-sample is $\text{Bias(T)} = \text{Bias(P)} + \text{Bias(E)}$, and the undercount estimate correcting for both sources of error is $\text{UC(T)} = \text{UC(DSE)} - \text{Bias(T)}$. Note that it is possible, as observed for EPS 2 and 4 in Table 12, that undercount estimate is negative, thus indicating an overcount instead. This happens when the DSE (or MLE) is less than CEN, the total census enumeration. The dual system data represents "corrected" census counts with erroneous and other incorrect enumerations excluded from CEN.

For each of Bias(P), Bias(E) and Bias(T), a positive estimate indicates an upward bias in the undercount estimate from the DSE by ignoring the corresponding source of error, that is, UC(DSE) should be reduced by the estimated bias to account for that source of error. For each of UC(DSE), UC(P), UC(E) and UC(T), we get significantly higher undercount figures for each of the five minority post-strata, *i.e.*, EPS 1, 3, 5, 8 and 11. For both Bias(P) and Bias(E), all the bias estimates are positive except for Bias(P) for post-stratum 9 and Bias(E) for post-stratum 11. This supports the common belief that there is usually an upward bias attributable to matching errors in the undercount estimate by the DSE, except for some non-minority geographical areas where in fact there is disproportionately large share of erroneous enumerations.

The effects of the two types of matching errors are well understood. False nonmatches results in upward bias and false matches produce downward bias. The nature of the overall matching bias is then dependent upon which type of matching error dominates. By computing undercount estimates for 1980 Census data with selective pair of γ and β , Ding (1990) concludes that due to high capture probabilities in the census application of the capture-recapture technique, the matching bias is dominated by the false nonmatch rate when the false nonmatch rate (γ) and the false match rate (β) are about the same magnitude. This point can be easily confirmed here. EPS 4 has the largest estimate of γ , $\hat{\gamma} = .021\%$ and results in the largest Bias(P) = 2.06%. EPS 3 and EPS 4 have about the same estimate of β , $\hat{\beta} = .012\%$ and $.013\%$, respectively, but EPS 3 has much smaller Bias(P) = .80%, due to smaller estimate of γ , $\hat{\gamma} = .010\%$. About a .01% difference in $\hat{\gamma}$ gives dramatic difference in Bias(P). For matches and nonmatches with complete data, Fay *et al.* (1988, p. 53) state "Because of sometimes difficult nature of the matching work, false nonmatches probably represent a greater concern than false matches". The data analyzed by our methods include both complete data and data produced as a result of the Bureau's imputation procedure. The sensitivity of our estimates to γ lends some support to the statement by Fay *et al.* when both matching for complete data and matching for imputed data are considered together. On the other hand, a downward bias can be observed when $\hat{\beta}$ is much larger than $\hat{\gamma}$. For EPS 9, $\hat{\beta} = .022\%$, about 10 times as large as $\hat{\gamma} = .002\%$. Thus false matches dominate false nonmatches for this stratum, and we see the only negative (downward) bias, Bias(P) = -.22%.

For a specific matching procedure there is an inevitable trade-off between matching errors and unresolved cases. Depending on the extent of unresolved cases and the imputation algorithm used, the resolution process might yield a significant number of false matches. The empirical evidence accumulated by the Bureau of the Census, as we note above, lends some support for the "unbiasedness"

of the missing data mechanism used in the imputation process in our example, but further evidence on the issue is desirable.

6. SUMMARY

In this article, we have presented models and methods for the estimation of population total and census undercount that corrects for matching bias of the usual dual-system estimate in the presence of matching errors. Two sources of information are combined in the estimation procedure, the dual-system or capture-recapture census data, and the data from a matching error study (rematch study). The accuracy of our estimates relies on the assumption that the rematch is error free. Matching error rates are likely not to be homogeneous over different population strata. Model (B') allows for heterogeneity of matching error rates across various population strata but requires stratified rematch data to estimate the error parameters within strata. The methods presented here generalize the standard theoretical framework for the use of maximum likelihood estimation to accommodate matching errors.

We can adjust for erroneous enumerations in the estimate of EE by the use of rematch data for the E-sample. We obtain an overall matching bias in the DSE by adding two bias components from the P-sample and the E-sample. Our analysis of the 1986 Los Angeles test census data indicates that the upward bias of the DSE in the estimate of the census undercount is just under 1%, thereby lending support to the 1% value used by Hogan and Wolter (1988) in their evaluation study. For the analysis on 1990 Census data, the computational results not only agree with understood aspects of matching bias, but also offer findings that were not previously known.

For simplicity, we have assumed that the PES is (allowing for stratification) based on simple random sampling. The models still need to be adapted to account for the complex sampling design actually used (see Hogan 1992, 1993).

It has been known that the perfect matching assumption does not hold in the application of dual system estimation in the U.S. census. The matching problem in the use of the DSE has two components. The first component involves the missing P-sample enumeration status. The second involves errors in classifying P-sample people as enumerated or not. The present paper provides a method to address both components using dual system data adjusted for imputed enumeration probabilities, and can be of possible value in future censuses provided that the models are adapted to handle the complex survey design of the PES. Ding (1993c) develops estimates to directly address the first component by modifying the usual DSE method and describes the relationship between the proposed estimates and those that result from the application of the Census Bureau's imputation scheme for missing P-sample enumeration status (Schenker 1988, Belin *et al.* 1993).

ACKNOWLEDGMENTS

Fienberg's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada to York University, Toronto, Canada. The authors are grateful to Mary Mulry for furnishing data on 1990 Decennial Census, to Joe Sedransk for suggestions, and to Jay Kadane, Larry Wasserman and Mike Meyer for commenting on an earlier version of this work. An Associate Editor and two referees provided comments that have led to a sharpening of the discussion. The basic models in this manuscript were first developed as part of the first author's Ph.D. thesis at Carnegie Mellon University.

APPENDIX

Comparison of Estimates from Weighted and Unweighted P-Sample Data

For simplicity, we assume a weight $k > 1$ for the P-sample and consider the usual dual system estimation problem. Let $\{x_{ij}\}$ be the cell counts in the 2×2 table for weighted P-sample data and census enumerations, $i, j = 1, 2$ and $ij \neq 22$. One could make inference with unweighted P-sample data and census enumerations deflated by a factor of k to get cell counts $\{y_{ij}\}$, $i, j = 1, 2$ and $ij \neq 22$. Then $x_{ij} = ky_{ij}$, $ij \neq 22$, and $x_{1+} = ky_{1+}$, $x_{+1} = ky_{+1}$. Let the usual dual system estimates derived from $\{x_{ij}\}$ be \hat{p}_1, \hat{p}_2 and \hat{N}_w , and estimates from $\{y_{ij}\}$ be \hat{q}_1, \hat{q}_2 and \hat{N}_u . The estimates are (Bishop *et al.* 1975, chap. 6) $\hat{p}_1 = x_{11}/x_{+1} = y_{11}/y_{+1} = \hat{q}_1$, $\hat{p}_2 = x_{11}/x_{1+} = y_{11}/y_{1+} = \hat{q}_2$, $\hat{N}_w = x_{1+}x_{+1}/x_{11} = ky_{1+}y_{+1}/y_{11} = k\hat{N}_u$. Thus if one considers the unweighted P-sample data and uses $\hat{N}^* = k\hat{N}_u$ to estimate the population total, then \hat{q}_1, \hat{q}_2 and \hat{N}^* give the same point estimates as \hat{p}_1, \hat{p}_2 and \hat{N}_w from weighted P-sample data. From the asymptotic normal distribution of the estimates (Ding 1993b), we have $\text{Var}(\hat{N}_w) = k\text{Var}(\hat{N}_u)$, $\text{Var}(\hat{q}_1) = k\text{Var}(\hat{p}_1)$, $\text{Var}(\hat{q}_2) = k\text{Var}(\hat{p}_2)$. Then $\text{Var}(\hat{N}^*) = k\text{Var}(\hat{N}_w)$, and \hat{q}_1, \hat{q}_2 and \hat{N}^* have larger variance than \hat{p}_1, \hat{p}_2 and \hat{N}_w , respectively. To compute estimates with unweighted P-sample data, one needs to know k and $\{y_{ij}\}$. We emphasize that the trivial case of a constant sampling weight for all cases in the same post-stratum is assumed here for simplicity of discussion. However, the real situation can be complex. For example, Blacks may be sampled at a low probability in a White stratum and are then combined with other Blacks sampled with much higher probabilities.

REFERENCES

- ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., and ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88, 1149-1166.
- BIEMER, P.P. (1988). Modeling matching error and its effect on estimates of census coverage error. *Survey Methodology*, 14, 117-134.
- BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press.
- CHAPMAN, D.C. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.
- CHILDERS, D., DIFFENDAL, G., HOGAN, H., and MULRY, M. (1989). Coverage Evaluation Research: the 1988 Dress Rehearsal. Paper presented to the Census Advisory Committee of the American Statistical Association and the Census Advisory Committee on Population Statistics at the Joint Advisory Committee Meeting, Alexandria, VA.
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of American Statistical Association*, 88, 1137-1148.
- DAVIS, M.C., MULRY, M., PARMER, R., and BIEMER, P. (1991). The matching error study for the 1990 Post Enumeration Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 248-253.
- DAVIS, M.C., and BIEMER, P. (1991a). Estimates of P-Sample Clerical Matching Error from a Rematching Evaluation. Report on Post-Enumeration Survey Evaluation Project P7, U.S. Department of Commerce, Bureau of the Census.
- DAVIS, M.C., and BIEMER, P. (1991b). Measurement of the Census Erroneous Enumerations: Clerical Error Made in the Assignment of Enumeration Status. Report on Post-Enumeration Survey Evaluation Project P10, U.S. Department of Commerce, Bureau of the Census.
- DING, Y. (1990). Capture-recapture Census with Uncertain Matching. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania.
- DING, Y. (1993a). On the asymptotic normality of multinomial population size estimates with application to the backcalculation estimates of AIDS epidemic. To appear in *Biometrika*.
- DING, Y. (1993b). On the asymptotic normality of dual system estimates. Unpublished manuscript.
- DING, Y. (1993c). Capture-recapture census with probabilistic matching. Submitted for publication.
- DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in central Los Angeles county. *Survey Methodology*, 14, 71-86.
- ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond (with discussion). *Journal of American Statistical Association*, 80, 98-131.

- FAY, R.E., PASSEL, J.S., ROBINSON, J.G., and COWAN, C.D. (1988). The Coverage of Population in the 1980 Census. U.S. Department of Commerce, Bureau of the Census.
- FIENBERG, S.E. (1989). Undercount in the U.S. decennial census. *Encyclopedia of Statistical Sciences, Supplement Volume*, 181-185.
- JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 test census of Tampa, Florida. *Journal of American Statistical Association*, 84, 414-420.
- JEFFERSON, T. (1986). Letter to David Humphreys. *The Papers of Thomas Jefferson*, 22, 62.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a post-enumeration survey. *Survey Methodology*, 14, 99-116.
- MULRY, M.H., and SPENCER, B.D. (1991). Total error in PES estimates of population: the dress rehearsal census of 1988 (with discussion). *Journal of American Statistical Association*, 86, 839-854.
- MULRY, M.H., and SPENCER, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of American Statistical Association*, 88, 1080-1091.
- RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.
- SANATHANAN, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.
- SCHENKER, N. (1988). Handling missing data in coverage estimation with application to the 1986 test of adjustment related operations. *Survey Methodology*, 14, 87-98.
- WOLTER, K. (1986). Some coverage error models for census data. *Journal of American Statistical Association*, 81, 338-346.

A Hypothesis Test of Linear Regression Coefficients with Survey Data

PHILLIP S. KOTT¹

ABSTRACT

This paper discusses testing a single hypothesis about linear regression coefficients based on sample survey data. It suggests that when the design-based linearization variance estimator for a regression coefficient is used it should be adjusted to reduce its slight model bias and that a Satterthwaite-like estimation of its effective degrees of freedom be made. A very important special case of this analysis is its application to domain means.

KEY WORDS: Design-based; Domain mean; Effective degrees of freedom; Model-dependent; Probability order.

1. INTRODUCTION

Most of statistical theory is analytical in nature. One begins with a set of data and a fairly general stochastic model believed to have generated that data. Statistical theory is then invoked to estimate the parameters of the model and to determine the accuracy of those estimates. Ultimately, the original model may be pared down as the result of a series of statistical tests which often take the form of investigations into whether particular parameter values may be reasonably inferred to be zero.

The bulk of survey sampling theory, by contrast, is not analytical but descriptive. There is a finite population of interest. Information about this population can, in principle, be summarized by means of one or more descriptive statistics (for example, the population mean and median). The survey statistician is constrained by time or budgetary considerations to estimate such statistics using only a sample of population units. He (she) often faces a two-fold problem: first a method of sample selection needs to be chosen, then the population statistic(s) needs to be estimated from the sample. Although it is possible to construct a model-dependent statistical theory for these purposes (see, for example, Royall 1970), most survey statisticians invoke a model-free approach known as design-based sampling theory. In this theory, it is not the sample data values that are stochastic (as they are in model-dependent theory) but the sample selection process. Rao and Bellhouse (1989) provides a useful summary of both design-based and model-dependent theory and of attempts to synthesize the two approaches.

The main concern here will be in the testing of a single hypothesis about linear regression parameters. We will assume that the model is correct and that model errors are normally distributed with a possibly complex covariance structure. Unlike Wu *et al.* (1988), we will not explicitly model the error structure (except, perhaps, at a latter

stage). Rather, we will focus our attention on a *t*-statistic calculated using the linearization variance estimator. That this variance estimator has desirable robustness properties from a model-dependent point of view has been demonstrated by Skinner (1989) and Kott (1991).

This paper will provide methods for reducing the model bias of the linearization variance estimator and for determining its effective degrees of freedom. A very important special case of this analysis is its application to the estimated variance of domain means and the difference of such means. Since the analysis in this paper is strictly model-dependent, the terms “bias” and “variance” will refer to model bias and model variance unless otherwise specified.

2. THE MODEL

Suppose we have a population of M elements that can be fit by the linear model:

$$y_M = X_M \beta + \epsilon_M, \quad (1)$$

where y_M is an $M \times 1$ vector of population values for the designated dependent variable;

X_M is an $M \times K$ matrix of population values for the K designated independent variables;

β is a $K \times 1$ vector of regression coefficients; and

ϵ_M is a normally distributed random vector with mean 0_M and variance Σ_M .

A random sample, S , of m distinct elements is drawn from the population. To allow a certain amount of generality in the sampling design, we assume that the population is divided into L strata. From each stratum h , n_h distinct clusters of elements are randomly sampled and denoted $u_{h1}, u_{h2}, \dots, u_{hn_h}$. A random sample of m_{hj} elements is selected from each cluster hj . The clusters are also referred to as primary sampling units. There are $n = \sum n_h$ primary sampling units in the sample.

¹ Phillip S. Kott, National Agricultural Statistics Service, 3201 Old Lee Highway, Fairfax, VA 22030, U.S.A.

Each sampled element has a designation hji , where h is its stratum, hj its primary sampling unit within h , and i the element itself within hj . Let p_{hji} be the probability that element hji is in the sample, and let $w_{hji} = m / (Mp_{hji})$ be the sampling weight of the element. Observe that the sampling weights have been normalized so that if p_{hji} equals the sampling fraction, m/M , then w_{hji} would be unity.

The linear model in (1) also applies to the elements in sample S :

$$y_S = X_S \beta + \epsilon_S,$$

where y_S , for example, is the $m \times 1$ vector of sampled values for the dependent variable. Let $\epsilon_{hj} = (\epsilon_{hj1}, \epsilon_{hj2}, \dots, \epsilon_{hjm_{hj}})$ be the error vector for the elements in primary sampling unit hj . Now, ϵ_S can be arranged so that the ϵ_{hj} are stacked one on top of the other. Let $\text{Var}(\epsilon_{hj}) = E(\epsilon_{hj}\epsilon_{hj}')$ be denoted by the $m_{hj} \times m_{hj}$ matrix Σ_{hj} , which need not be diagonal. We assume that the ϵ_{hj} are uncorrelated across primary sampling units, so that Σ_S is block diagonal.

The design-based estimator for β is the weighted least squares estimator:

$$b_W = (X_S' W X_S)^{-1} X_S' W y_S,$$

where W is the $m \times m$ diagonal matrix of sampling weights. The g -th diagonal value of W is the sampling weight associated with the g -th element of the sample. Clearly, b_W is an unbiased estimator of β under the model in (1).

One can simplify the notation for b_W by letting C be the $k \times m$ matrix $(X_S' W X_S)^{-1} X_S' W$, so that $b_W = C y_S$. Let D_{hj} be a $m \times m$ diagonal matrix with 1's corresponding to the sampled elements of hj and 0's elsewhere. Furthermore, let $C_{hj} = C D_{hj}$. Finally, let $r_S = y_S - X_S b_W$ be the vector of residuals.

The Taylor series or linearization estimator for the mean squared error of b_W (Shah *et al.* 1977) is

$$\text{mse} = \sum_{h=1}^L (n_h / [n_h - 1]) \sum_{j=1}^{n_h} A_{hj} r_S r_S' A_{hj}', \quad (2)$$

where $A_{hj} = C_{hj} - n_h^{-1} \sum C_{hg}$, and the summation is over all the primary sampling units in stratum h . The terms "Taylor series" and "linearization" refer to the derivation of mse using design-based sampling theory. Kott (1991) shows that mse is a nearly unbiased estimator of the model variance of b_W under reasonable conditions.

It should be noted that in their derivation of mse, Shah *et al.* assumed that the primary sampling units were chosen with replacement. Here, as in Kott (1991), we are assuming that the primary sampling units are distinct which suggests that they were selected *without* replacement. The reason

for this discrepancy is that the assurance of independence among the selected primary sampling units within a stratum in design-based theory and model-dependent theory has almost opposite requirements. The discrepancy goes away, however, if we assume that the primary sampling units were chosen without replacement but that the goal of design-based regression theory is not to estimate a finite population regression parameter but the limit of that parameter as the population (and the number of primary sampling units per stratum) grows arbitrarily large. See Fuller (1975).

If the model in equation (1) holds and $L > 1$, then there is an alternative to mse that is also nearly unbiased. It has the same form as equation (2) except that all n sampled primary sampling units are treated as if they came from a single stratum ($L = 1$). Since the alternative can be expressed using equation (2), there is no need to treat it separately in the analysis that follows.

3. A CONVENTIONAL DESIGN-BASED t -STATISTIC

The estimator b_W is a K -vector. In this section we will be interested in the t -statistic used to test the univariate hypothesis that $q\beta = \Theta_0$ for some K element row vector $q = (q_1, q_2, \dots, q_K)$. The most common example of such an hypothesis addresses whether a particular element of $\beta = (\beta_1, \dots, \beta_K)$, say β_k , is zero. In this example, all of the q_i would be zero except q_k which would be 1; Θ_0 would also be zero.

If the model in (1) and the null hypothesis that $q\beta = \Theta_0$ are true, then

$$\Theta = (qb_W - \Theta_0) / \{q \text{Var}(b_W) q'\}^{1/2}$$

would be normally distributed with mean 0 and variance 1. If $\text{Var}(b_W)$ were known, the null hypothesis could be tested by comparing the statistic Θ to a standard normal table. Unfortunately, $\text{Var}(b_W)$ must be estimated from the sample. Conventional design-based practice is to compare the statistic

$$t = (qb_W - \Theta_0) / (qmseq')^{1/2}, \quad (3)$$

to a Student's t distribution with $n - L$ or $(n - L - K)$ degrees of freedom (see Shah *et al.* 1977).

The primary goal of this paper is to investigate and then modify the rather *ad hoc* practice described above using the model in equation (1) and our assumptions that Σ_S is block-diagonal. This will be done by investigating $s^2 = qmseq'$ as an estimator for $v^2 = q \text{Var}(b_W) q'$. First, s^2 will be adjusted to reduce its bias; then, a better determination of the adjusted estimator's effective degrees of freedom will be established.

4. THE MODEL BIAS OF s^2

The analysis to be conducted is asymptotic. Many of the results rely on the assumption that n , the number of primary sampling units in the sample, is large. (Formally, we should assume that there are infinite sequences of statistics taking on values as n grows arbitrarily large.) If n is large, then so too must be M and m , the number of elements in the population and the sample, respectively. We will assume that $\max\{m_{hj}\}$ is bounded by a finite value, say \bar{m}_0 . Thus, m is bounded by $\bar{m}_0 n$ and the number of nonzero elements in the block-diagonal matrix Σ_S is bounded by $\bar{m}_0^2 n$.

The number of columns of X_S , K , is assumed to be fixed, but we have some flexibility concerning the number of strata, L . Either L can stay fixed as n grows arbitrarily large with the n_h/n ratios converging to fixed positive limits, or L/n can converge to a fixed positive limit with $\max\{n_h\}$ bounded.

Our concern here is with providing *sufficient* conditions for the subsequent analysis in the text to hold. The random variable ϕ (formally, the infinite random sequence $\{\phi_n\}$) will be said to be of probability order $n^{-\delta}$, i.e., $\phi = O_P(n^{-\delta})$, when $|E(\phi^2)| < B/n^{2\delta}$ for some finite B . Similarly, the random matrix Φ will be said to equal $O_P(n^{-\delta})$ when each element ϕ_{ij} in Φ satisfies $|E(\phi_{ij}^2)| < B/n^{2\delta}$. When ϕ is not random, the P subscript on O is not needed. The same is true for O .

The following assumptions are reasonable given the structure that has been laid out:

- (1) $C = (X'WX)^{-1}X'W$ exists and is $O(1/n)$, and
- (2) $E(\hat{\Sigma}_{hj}) = \Sigma_{hj} + O(1/n)$, where $\hat{\Sigma}_{hj} = r_{hj}r'_{hj}$.

Assumption 1 assures us that $\text{Var}(b_W) = C\Sigma_S C' = O(1/n)$ since there are m elements in the rows of C and no more than $\bar{m}_0^2 n$ non-zero elements in Σ_S .

The variance of qb_W can be rewritten as $v^2 = \sum \sum v_{hj}/n^2$, where $v_{hj} = n^2 g_{hj} \Sigma_S g_{hj}$, $g_{hj} = qCD_{hj}$, and D_{hj} is a diagonal matrix with 1's corresponding to the sampled elements of primary sampling unit hj and 0's elsewhere. Similarly, $s^2 = qmseq'$ can be rewritten as

$$\begin{aligned} s^2 &= \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (g_{hj} - g_h) r_S r'_S (g_{hj} - g_h)' \\ &= \sum (n_h/[n_h - 1]) \sum [g_{hj} \hat{\Sigma}_S g'_{hj} \\ &\quad - 2g_h \hat{\Sigma}_S g'_{hj} + g_h \hat{\Sigma}_S g'_h], \end{aligned} \quad (4)$$

where $g_h = \sum g_{hj}/n_h$, the summation is across the j in h , and $\hat{\Sigma}_S = \sum \sum D_{hj} r_S r'_S D_{hj}$.

Both g_{hj} and g_h are $O(1/n)$ because $C = O(1/n)$ and D_{hj} has a bounded number of non-zero values. Thus,

$$E(g_{hj} \hat{\Sigma}_S g'_{hj}) = g_{hj} \Sigma_S g'_{hj} + O(n^{-3}), \quad E(g_h \hat{\Sigma}_S g'_h) = g_h \Sigma_S g'_h + O(n^{-3}), \quad \text{and} \quad E(g_h \hat{\Sigma}_S g'_h) = g_h \Sigma_S g'_h + O(n^{-3}).$$

Consequently, $E(s^2 - v^2) = O(n^{-2})$.

Since $r_S = (I_m - XC)\epsilon_S$ and $E(\epsilon_S \epsilon'_S) = \Sigma_S$, $E(r_S r'_S) = \Sigma_S - XC\Sigma_S - \Sigma_S C'X' + XC\Sigma_S C'X'$. From equation (4), we can see that $E(s^2) = v^2 - R$, where $R = \sum (n_h/[n_h - 1]) \sum (g_{hj} - g_h) Z (g_{hj} - g_h)'$ and $Z = 2XC\Sigma_S - XC\Sigma_S C'X'$. Now $Z = O(1/n)$, because $C = O(1/n)$, X has a fixed number of columns, and the number of non-zero terms in any column of Σ_S is bounded. This implies $R = O(n^{-2})$. Thus, $-R/v^2$, the relative bias of s^2 , is $O(1/n)$.

An alternative estimator for v^2 with a reduced relative bias is

$$s_*^2 = s^2 / (1 - s^{-2} \hat{R}), \quad (5)$$

where

$$\hat{R} = \left\{ \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (g_{hj} - g_h) \hat{Z} (g_{hj} - g_h)' \right\},$$

and

$$\hat{Z} = 2XC\hat{\Sigma}_S - XC\hat{\Sigma}_S C'X'.$$

In equation (5), \hat{R}/s^2 is used to estimate R/v^2 . The variance estimator s_*^2 has been proposed here rather than the more obvious $s^2 + \hat{R}$ as *ad hoc* compensation for the slight relative bias of \hat{R} as an estimator of R .

5. THE RELATIVE VARIANCE OF THE VARIANCE ESTIMATOR

Let $e_{hj} = ng_{hj}\epsilon_S$ so that $\text{Var}(e_{hj}) = v_{hj}^2$, and recall that $v^2 = \sum \sum v_{hj}^2/n^2$. If $\hat{e}_{hj} = ng_{hj}r_S$, then the random variable s^2 can be re-written as

$$\begin{aligned} s^2 &= n^{-2} \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (\hat{e}_{hj} - \hat{e}_h)^2 \\ &= n^{-2} \sum (n_h/[n_h - 1]) \{ \sum (e_{hj} - e_h)^2 \\ &\quad - (g_{hj} - g_h) A (g_{hj} - g_h)' \}, \end{aligned}$$

where $A = 2XCe_S e'_S - XCe_S e'_S C'X'$. It is now possible to show that

$$\begin{aligned} s_*^2 &= n^{-2} \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{j=1}^{n_h} (e_{hj} - e_h)^2 \\ &\quad + O_P(n^{-5/2}). \end{aligned}$$

Consider a random variable with a χ^2 distribution with F degrees of freedom. Its relative variance is $2/F$. This suggests a Satterthwaite-like determination of the effective degrees of freedom of s_*^2 (see Satterthwaite 1946); namely,

$$F = \frac{(nv)^4}{\sum_{i=1}^L \left\{ \sum_{j=1}^{n_h} v_{hj}^4 + \sum_{j' \neq j} v_{hj}^2 v_{hj'}^2 / (n_h - 1)^2 \right\}}, \quad (6)$$

which is approximately 2 divided by the relative variance of s_*^2 (since $s_*^2 \approx n^{-2} \sum_h \{ \sum_j e_{hj}^2 + \sum_{j' \neq j} e_{hj} e_{hj'} / (n_h - 1) \}$).

What is being recommended here is that one tests whether $q\beta = \Theta_0$ by assuming under the null hypothesis that

$$t_* = (qb_W - \Theta_0)/s_*, \quad (7)$$

has a Student's t distribution with F degrees of freedom, where F is determined using equation (6) and making some assumptions about the v_{hj} . Let us call this test the *adjusted t-test*.

6. A SIMPLE EXAMPLE

Consider a simple random sample of n units, n_1 of which are in a subset of the sample denoted by A and n_2 in the complement denoted \bar{A} . Let y_i be the observed value for unit i . Suppose the following linear model holds:

$$y_i = d_i \beta_1 + (1 - d_i) \beta_2 + \epsilon_i, \quad (8)$$

where $d_i = 1$ if unit i is in set A , and 0 if i is in \bar{A} ; and the ϵ_i are independent normally distributed random variables.

Assuming homoscedastic errors, both the model-dependent and design-based regression estimator for β_1 is the simple domain mean, $\bar{y}_A = \sum_{i \in A} y_i / n_1$. The linearization estimator for the variance of this estimator is simply $v_L = (n/[n-1]) \sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1^2$. (It should be noted that when a domain mean is viewed as an analytic parameter, its variance requires no finite population correction; see Fuller 1975).

This linearization estimator, v_L , differs from model-dependent variance estimator: $v_M = [\sum_{i \in A} (y_i - \bar{y}_A)^2 + \sum_{i \in \bar{A}} (y_i - \bar{y}_A)^2] / [n_1(n-2)]$. The advantage of v_L is that, unlike v_M , it is asymptotically unbiased under the model even when the ϵ_i are heteroscedastic. This point was noted by Skinner (1989) and Kott (1991). Unfortunately, there still may be considerable bias for finite n . For example, when $n = 100$ and $n_1 = 10$, the relative bias of v_L is approximately 10%. We can see this by noting that $v_E = \sum_{i \in A} (y_i - \bar{y}_A)^2 / (n_1[n_1 - 1]) = ([n-1]/n) (n_1/[n_1 - 1]) v_L$ is exactly unbiased.

Continuing the example: If one were to calculate a t -statistic using conventional design-based practice, he (she) would not only use a biased variance estimator but would also assume that the statistic has 97 or 99 degrees of freedom (100 sampling units minus one strata minus two regressors, were this last subtraction is not always performed). Under ideal conditions (homoscedastic errors within set A), however, the t -statistic calculated using v_E has a Student's t distribution with only 9 degrees of freedom.

Applying equation (5) to the linearization variance estimator, v_L , produces a variance estimator virtually identical to v_E (since $\hat{R} = [v_L/n_1[1 - n_1/n]]$, s_*^2 differs from v_E by only 0.1%). Assuming identically distributed errors within sets A and \bar{A} , calculating the effective degrees of freedom, F , with equation (6) yields 9.99. This is almost exactly one degree too many but clearly better than 97 or 99.

A natural hypothesis to test is whether the domain means, β_1 and β_2 , in equation (8) are equal. In other words is $\beta_1 - \beta_2 = \Theta_0 = 0$? Assuming that all units have the same variance, the adjusted t statistic is

$$t_* = \frac{\sum_{i \in A} y_i / n_1 - \sum_{i \in \bar{A}} y_i / n_2}{(1 - s^{-2} \hat{R})^{1/2} s},$$

where

$$s^2 = [n/(n-1)] [\sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1^2 + \sum_{i \in \bar{A}} (y_i - \bar{y}_A)^2 / n_2^2],$$

and

$$\hat{R} = [n/(n-1)] [(\sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1^3) (1 - n_1/n) + (\sum_{i \in \bar{A}} (y_i - \bar{y}_A)^2 / n_2^3) (1 - n_2/n)].$$

To calculate the effective degrees of freedom for $ns^2/(n-1)$ - and thus t_* - using equation (6), note that $L = 1$, and $v_i \propto 1/n_1^2$ for $i \in A$ while $v_i \propto 1/n_2^2$ for $i \in \bar{A}$. As a result,

$$F = \frac{(1/n_1 + 1/n_2)^2}{(1/n_1^3 + 1/n_2^3 + [\{1/n_1 + 1/n_2\}^2 - 1/n_1 - 1/n_2]/n^2)},$$

which is 12.3 when $n_1 = 10$ and $n_2 = 90$. The actual degrees of $ns^2/(n-1)$ (i.e., 2 divided by its relative variance) is reasonably close, 11.1 (the relative variance of $ns^2/(n-1)$ is $2[(n_1-1)/n_1^4 + (n_2-1)/n_2^4]$ divided by $[(n_1-1)/n_1^2 + (n_2-1)/n_2^2]^2$).

What this synthetic example principally shows is how misleading conventional design-based practice can be even with an apparently large sample size. The adjusted t -test is clearly a giant step in the right direction.

It is tempting to try to avoid making an assumption about the v_{hj} and to estimate F with

$$f = \frac{(ns_0)^4 - \sum_{j=1}^L \sum_{h_j}^{n_h} 2s_{hj}^4/3}{\sum_{i=1}^L \left\{ \sum_{j=1}^{n_h} s_{hj}^4/3 + \sum_{j' \neq j} s_{hj}^2 s_{hj'}^2 / (n_h - 1)^2 \right\}}, \quad (9)$$

where $s_{hj}^2 = n^2 (g_{hj} r_s)^2$. Although f is a consistent estimator of F , its use can produce misleading results as we shall see.

Repeated application of equation (9) on 10,000 simulated data sets constructed under the assumption that the ϵ_i in equation (8) are normal, independent, and identically distributed yielded an average f value for the variance of \bar{y}_A of approximately 11.2 with a standard deviation of about 3.5. In addition to its variability, the average f value is greater than F . This is due to the denominator of equation (9) itself being a random variable. It happens that the value of $1/f$ is roughly 0.100 ($\approx 1/9.99$), as expected. Thus, even though the use of f in equation (9) may seem appealing, it is not recommended.

7. ANOTHER EXAMPLE OF A DOMAIN MEAN

Faced with the simple example of the last section, most design-based statisticians would simply treat the units sampled from set A as an independent simple random sample. The linearization and model-dependent variance estimator would then coincide. In practice, however, samples often involve clustering, stratification, and unequal probabilities of selection. When the domain of interest is not a design stratum, it usually becomes impossible to separate out the domain's sampled elements (which need not be primary sampling units) and treat them as an independent random sample.

An example of such a complex sample is the 1985 Continuing Survey of Intakes by Individuals (CSFII). This was a stratified, multistage survey of the dietary intakes of women from 19 to 50 years of age and children from 1 to 5. There were roughly 140 women in the sample who described themselves as black and 1,150 who described themselves as white.

Assuming that a dietary intake value for each individual was independent and identically distributed, values of the relative variance of the linearization variance estimator (R/s^2 from equation (5)) and its effective degrees of freedom (F from equation (6)) were calculated for the two

race domains. The relative bias for white women was .003, while the effective degrees of freedom were 48.1. For black women, the relative bias was 0.026, and the effective degrees of freedom 10.1. Thus, even with a fairly large sample size, the effective number of degrees of freedom for black women was relatively small. The conventional determination of degrees of freedom was around 60 (120 PSU's minus 60 design strata).

8. DISCUSSION

As pointed out earlier, the use of design-based techniques can often provide protection when the model in equation (1) fails. Unfortunately, this protection can not be addressed in the strictly model-dependent framework adopted here. It would be unrealistic, however, to expect a conventional design-based t -statistic to behave any better when the model in equation (1) fails than when it holds.

One potential problem of the modified design-based test statistic suggested here occurs when the model in equation (1) does *not* fail: it may not be very powerful. Power can be lost by estimating regression coefficients with sampling weights and by not modelling the error structure directly.

This loss of power is due to the original design-based formulation and not to our modification of it. In fact, s_w^2 is a design consistent estimator of the design mean squared error of b_W whenever s^2 is. This is because \hat{R}/s^2 in equation (5) is also $O_P(1/n)$ from a design-based point of view assuming that the first stage of sampling is conducted *with* replacement.

Returning to the simple example of Section 6 can illustrate the issue of power forcefully. The model-dependent and design-based estimates are the same. If all the ϵ_i are assumed to be identically distributed, then the model-dependent variance estimator, v_M , which depends on the assumption of homoscedasticity, is unbiased and has 98 degrees of freedom. The adjusted design based variance estimator is also virtually unbiased, but it has only 9 degrees of freedom.

Often in practice, it will be prudent to sacrifice power for robustness. When that is the case, equation (6) provides an attractive method of measuring how much power may be lost using a modified design-based t -test (equation (7)) when the assumptions of the model are, in fact, correct. Furthermore, the equation lends itself to sensitivity analyses in which the effects of alternative assumptions about the v_{hj} can be evaluated.

ACKNOWLEDGEMENTS

The author would like to thank the staff of the Beltsville Human Nutrition Research Center for its support of this research and an associate editor and his (her) referees for their helpful comments.

REFERENCES

- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā, C*, 37, 117-132.
- KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, 107-112.
- RAO, J.N.K., and BELLHOUSE, D.R. (1989). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *American Statistical Association Proceedings Sesquicentennial Invited Paper Sessions*, 406-428.
- ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- SATTERTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- SHAH, B.V., HOLT, M.M., and FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 47, 43-57.
- SKINNER, C.J. (1989). Domain means, regression, and multivariate analysis. In *Analysis of Complex Surveys* (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley, 59-88.
- WU, C.J.F., HOLT, D., and HOLMES, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-159.

Matrix Masking Methods for Disclosure Limitation in Microdata

LAWRENCE H. COX¹

ABSTRACT

The statistical literature contains many methods for disclosure limitation in microdata. However, their use by statistical agencies and understanding of their properties and effects has been limited. For purposes of furthering research and use of these methods, and facilitating their evaluation and quality assurance, it would be desirable to formulate them within a single framework. A framework called *matrix masking* – based on ordinary matrix arithmetic – is presented, and explicit matrix mask formulations are given for the principal microdata disclosure limitation methods in current use. This enables improved understanding and implementation of these methods by statistical agencies and other practitioners.

KEY WORDS: Statistical confidentiality; Survey data processing; Mathematical methods.

1. INTRODUCTION

In this Information Age critical activities of society are fuelled by data. Users of statistical data rely especially upon government statistical agencies to collect reliable data and disseminate it in a timely and broadly useful way. Prior to the 1950s, data were released only in printed, tabulated form. Beginning in the 1960s, data at the individual respondent level – *statistical microdata* – began to be released by the U.S. Government.

At present, use of microdata outside statistical agencies for research and policy analysis is often curtailed because appropriate data are not released to users due to confidentiality concerns. For three decades statistical agencies have wrestled with policy and technical issues in microdata release, many of which remain unresolved (Federal Committee on Statistical Methodology 1994). The purpose of this article is to present a class of matrix transformations of microdata intended to help deal with this issue.

Duncan (1990) and Duncan and Pearson (1991) characterized several disclosure limitation methods for microdata – *microdata masks* – by means of matrix addition and multiplication, and named such characterizations “matrix masks.” Cox (1991) generalized the concept of matrix masks, and extended the characterization to other microdata masks. The characterization of microdata masks as matrix masks offers conceptual and statistical advantages. Matrix masking provides a simple language to represent, compare and evaluate microdata masking methods. Matrix masking expresses complicated, diverse methods in a form presentable to a wide audience including statisticians and data users, and offers a standard format to develop and optimize the efficiency of transportable microdata masking software.

In this paper, the concept of matrix masks is developed in a mathematically rigorous way. Explicit matrix mask formulations are provided for the principal microdata masking methods in current use, extending those presented in Duncan and Pearson (1991) and Cox (1991). This enables straightforward implementation of these methods in software, and facilitates closer examination and use of microdata masks by statistical agencies. This should lead to improved understanding of the properties of microdata masks and much needed understanding of their effects on data use.

2. MATRIX MASKS

2.1 Definitions

A microdata file containing p attribute values for each of n (respondent-level) data records can be represented as an $n \times p$ matrix X whose entries are denoted x_{ij} . Unless stated otherwise, X contains no missing values. A *matrix mask* (A, B, C) is a transformation of X of the form: $\tilde{X} = AXB + C$, with $A, B \neq 0$, involving ordinary matrix addition and multiplication. As A operates across the rows of X , A is called a *record transforming mask*. B is an *attribute transforming mask*, and C is a *displacing mask* (Duncan and Pearson 1991).

An *elementary matrix mask* of X is a matrix mask of the form AX , XB , or $X + C$. Iterations of (elementary) matrix masks of X are also matrix masks of X . Therefore, a matrix mask of X has the form $\tilde{X} = A\hat{X}B + C$, where either $\hat{X} = X$ or \hat{X} has been obtained from X by application of a sequence of elementary matrix masks. An important advantage of this definition is to enable different statistical disclosure limitation methods to be applied selectively to arbitrary subsets of the records and attributes of X (Section 4).

¹ Lawrence H. Cox, Senior Statistician, U.S. Environmental Protection Agency, AREAL (MD-75), Research Triangle Park, NC 27711, U.S.A.

The matrices A , B , C are not necessarily fixed. For example, a common mask for numeric attributes involves addition of random noise (Tendick 1991), so that C is a random matrix. The matrices A , B , C may depend upon X . For example, to displace X by additive random noise proportional to size, draw the c_{ij} randomly from a normal distribution with mean zero and standard deviation a multiple of $|x_{ij}|$, and set $\tilde{X} = X + C$. Or, with $A = X'$, $M = AX$ is sufficient for ordinary least squares regression (Duncan and Pearson 1991).

2.2 Notation

I denotes the identity matrix. Z denotes the matrix all of whose entries are zero, and J the matrix of all ones. U_{ij} denotes the matrix all of whose entries equal zero, except $u_{ij} = 1$. I is always a square matrix; Z , J and U_{ij} need not be. The U_{ij} matrix, when used as a pre-(post-)multiplier retains the values of only one row (column) of the matrix it multiplies. The dimensions of submatrices may vary between or within individual formulations and will be specified for clarity.

3. REPRESENTATIONS OF DATA MASKS AS ELEMENTARY MATRIX MASKS

3.1 Removing and Selecting Microdata

The most intuitively obvious method for limiting disclosure is to withhold certain microdata from release to data users. Typically, these data are associated with the highest disclosure risk and may require suppressing attributes (columns) or suppressing records (rows) of X prior to release.

Attribute suppression of the k -th attribute can be represented as an attribute transforming mask $\tilde{X} = XB$, where B is the $p \times (p - 1)$ block matrix:

$$\text{Supp}(k) = \begin{bmatrix} I & Z \\ Z & I \end{bmatrix},$$

whose upper I -matrix is of dimension $(k - 1) \times (k - 1)$, whose lower I -matrix is of dimension $(p - k) \times (p - k)$, and whose central Z -matrix is of dimension $1 \times (p - 1)$. An alternative formulation is $\text{Supp}(k) = \sum_{j < k} U_{jj} + \sum_{j > k} U_{j,j-1}$.

Suppression of several attributes can be represented as a product of B -matrices of this form. For example, $\text{Supp}(k)\text{Supp}(j)$ first suppresses the k -th attribute of X , and then suppresses the j -th attribute of the resulting $n \times (p - 1)$ dimensional matrix $X\text{Supp}(k)$. The dimensions of $\text{Supp}(k)$ and $\text{Supp}(j)$ are $p \times (p - 1)$ and $(p - 1) \times (p - 2)$.

It is sometimes necessary to delete individual records from X . For example, a respondent may have high identification risk, or a record may be out of scope or spurious. *Record deletion* of the h -th record can be represented as a record transforming mask $\tilde{X} = AX$, where A is an $(n - 1) \times n$ dimensional block matrix identical in structure to $\text{Supp}(h)$, except: the central Z -matrix of A is of dimension $(n - 1) \times 1$ and the dimensions of the upper and lower I -matrices of A are $(h - 1) \times (h - 1)$ and $(n - h) \times (n - h)$. This A -matrix is denoted $\text{Del}(h)$. An alternative formulation is $\text{Del}(h) = \sum_{i < h} U_{ii} + \sum_{i > h} U_{i-1,i}$.

Deletion of more than one record is represented as a product of A -matrices $\text{Del}(h)$. For example, to delete the h -th and i -th records of X , with $i > h$, use $\text{Del}(i - 1)\text{Del}(h)$. For $i < h$, use $\text{Del}(i)\text{Del}(h)$. The dimensions of $\text{Del}(i - 1)$ and $\text{Del}(h)$ are $(n - 2) \times (n - 1)$ and $(n - 1) \times n$.

The A -matrix that *systematically deletes* every h -th record (for $n = rh$; r an integer) is a block matrix comprising r vertical blocks $\text{Del}(h)$, each of dimension $(h - 1) \times n$. This generalizes to nonsystematic deletion.

The complement of record deletion is *record sampling*. The A -matrix that systematically samples every h -th record of X , for $n = rh$, is an $r \times n$ matrix whose q -th row is the $1 \times n$ dimensional U -matrix $U_{1,qh}$. More generally, to draw a sample of size s comprising the records of X indexed by the set $S = \{s_v : v = 1, \dots, s\}$, use the A -matrix $\text{Sam}(X, S)$ of dimension $s \times n$, each row of which is a U -matrix U_{1,s_v} of dimension $1 \times n$.

3.2 Aggregating and Grouping Microdata

The risk of a respondent being identified and confidential data disclosed tends to decrease as data are more highly aggregated. *Attribute aggregation* and other microdata masks are based on this principle.

The aggregation mask that replaces the first of two attributes (the j -th attribute) by the sum of the two attributes, and deletes the second attribute (the k -th attribute) from X , for $j < k$, can be represented as an attribute transformation $\tilde{X} = XB$, where B is the $p \times (p - 1)$ dimensional block matrix:

$$\text{Agg}(j,k) = \begin{bmatrix} I & Z \\ U_{1j} & \\ Z & I \end{bmatrix}.$$

The upper I -matrix of $\text{Agg}(j,k)$ is of dimension $(k - 1) \times (k - 1)$, the lower I -matrix is of dimension $(p - k) \times (p - k)$, and the central U -matrix U_{1j} is of dimension $1 \times (p - 1)$. Alternative formulations are

$$\text{Agg}(j,k) = \text{Supp}(k) + U_{kj}, \quad \text{for } j < k, \quad \text{and}$$

$$\text{Agg}(j,k) = \text{Supp}(k) + U_{k,j-1}, \quad \text{for } j > k.$$

Aggregation-deletion over more than two attributes can be represented as a product of \mathbf{B} -matrices of this form. Construct \mathbf{B}_1 as above to aggregate the first two attributes to a subtotal, replace the first attribute by the subtotal, and delete the second attribute. Proceed iteratively forming $\mathbf{B}_2, \dots, \mathbf{B}_{c-1}$ until all summand attributes have been incorporated into the total and deleted. Then $\mathbf{B} = \mathbf{B}_1 \cdots \mathbf{B}_{c-1}$.

An alternative formulation for aggregation of the j -th and k -th attributes, replacement of the j -th attribute, and deletion of the k -th attribute, is given by the \mathbf{B} -matrix product $\mathbf{Add}(j, k) \mathbf{Supp}(k)$. Aggregation and replacement of the j -th attribute without deleting the k -th attribute can be accomplished using the $p \times p$ dimensional \mathbf{B} -matrix: $\mathbf{Add}(j, k) = \mathbf{I} + \mathbf{U}_{kj}$. This generalizes to more summands v by adding more \mathbf{U}_{vj} . To create a new totals attribute (attribute $p + 1$) from the j -th and k -th attributes without replacing either attribute, form the $p \times (p + 1)$ dimensional \mathbf{B} -matrix $[\mathbf{I} \mid \mathbf{U}_{j1} + \mathbf{U}_{k1}]$, whose \mathbf{I} -matrix is of dimension $p \times p$, and whose right-hand submatrix is of dimension $p \times 1$. Aggregating another attribute v amounts to adding additional \mathbf{U}_{v1} to the right-hand submatrix.

Grouping categorical data, sometimes referred to as *collapsing categories*, is representable as attribute aggregation. Represent each of the c mutually exclusive categories of a categorical variable by a column of \mathbf{X} . The absence (presence) of the corresponding trait is represented in each column by 0 (1). Grouping the c attribute categories to form one combined category is simply aggregation across the c attributes, replacing one attribute by the aggregate and deleting the remaining attributes, using \mathbf{B} -matrices in the manner described above.

It is sometimes desirable to aggregate attribute values across microrecords. For example, if microrecords can be grouped according to some notion of "similarity" (e.g., age or profession, or total value of shipments or size of work force for businesses in a particular industry), then an alternative to releasing high risk microrecords is to release a microdata file whose records are *microaggregates* or *microaverages* of subsets of the original records.

Record aggregation can be performed in several ways. A typical case is to replace all summands by the corresponding totals. Assume that the records to be microaggregated are arranged consecutively, and denote the respective sizes of the record groups by n_1, n_2, \dots, n_s , where $n = n_1 + n_2 + \dots + n_s$. Microaggregation can be accomplished using a diagonal block \mathbf{A} -matrix of dimension $n \times n$. The main diagonal of \mathbf{A} is comprised of an ordered block of square \mathbf{J} -matrices of dimension $n_v \times n_v$, $v = 1, \dots, s$; the remaining entries of \mathbf{A} are zero. Under microaggregation (microaveraging), original values are replaced by microaggregates (microaverages) in each record of the aggregation group. Alternatively, in each group one record may be replaced by the microaggregated record while the other records are deleted. This may be

accomplished using \mathbf{J} -matrices of dimension $1 \times n_v$, in which case the dimension of \mathbf{A} is $s \times n$. To construct microaverages in lieu of microaggregates, each \mathbf{J} -matrix is replaced by its corresponding $(1/n_v)\mathbf{J}$.

3.3 Scrambling Record Order

A microdata file \mathbf{X} being prepared for public use is typically derived from a larger data file (e.g., by sampling) or from a more detailed file (e.g., by removal of directly identifying information such as name, address, and social security number). The larger file is often maintained in a prescribed sort order, such as by geography or social security number, and \mathbf{X} is apt to inherit this ordering. To reduce disclosure risk, the order of the microrecords of \mathbf{X} must be *scrambled*. Record scrambling can be accomplished using a stochastic \mathbf{A} -matrix. Given a reordering of the rows (records) of \mathbf{X} (i.e., a permutation \mathbf{P} of the row numbers $\{1, \dots, n\}$), then for $\mathbf{P}(i) = h$, set the i -th row of \mathbf{A} equal to the \mathbf{U} -matrix \mathbf{U}_{1h} of dimension $1 \times n$. \mathbf{A} is denoted $\mathbf{Reo}(\mathbf{P})$. An alternative formulation is $\mathbf{Reo}(\mathbf{P}) = \sum_{i=1}^n \mathbf{U}_{i, \mathbf{P}(i)}$.

3.4 Rounding and Perturbing Microdata

Data rounding is used by statistical agencies for several purposes, including disclosure limitation. Integer variables such as age or years worked, or number of children, presented exactly, could be used in combination with other information to identify respondents (Bethlehem, Keller and Pannekoek 1990). *Conventional rounding* (e.g., base 5, remainders 0, 1, 2 are rounded down; remainders of 3, 4 are rounded up), does not preserve additivity to totals, and *controlled rounding*, designed to preserve additivity to totals in one and two way tabulations, may be preferred (Cox and Ernst 1982). Methods are also available for *unbiased controlled rounding* in one- or two-way tables (Cox 1987).

Data perturbation limits disclosure by introducing slight changes to microdata values. Additive perturbation amounts to adding appropriate perturbation values to original values. Additive perturbation values are often drawn randomly from a distribution with mean zero and variance small relative to that of the data. Nonrandom perturbation is also used.

Rounding and additive perturbation can be represented as displacing masks. For each value x_{ij} , the displacement c_{ij} to x_{ij} is computed according to the rounding or perturbation algorithm, with $c_{ij} = 0$ for those values not subject to change. Then, $\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{C}$ is the matrix of rounded (perturbed) values.

3.5 Attribute Topcoding

Attribute topcoding is a method by which, given a predetermined (large) value T_j of the j -th attribute, all values $x_{ij} > T_j$ are replaced by T_j . Given $x_{ij} = f_{ij} T_j + r_{ij}$,

for f_{ij} the integer quotient, and r_{ij} the remainder, $0 \leq r_{ij} < T_j$, compute $t_{ij} = (\text{Max}\{r_{ij}, (T_j + 1)^{f_{ij}} - 1\}) \bmod (T_j + 1)$. To topcode X , use the displacing mask $\text{Tco}(X) = (t_{ij} - x_{ij})$.

4. REPRESENTATIONS OF DATA MASKS AS MATRIX MASKS

4.1 Selecting and Modifying Attribute-Record Combinations

The formulations of the preceding section, based on elementary matrix masks, are applied to the entire microdata file X , and do not enable selective masking of arbitrary subsets of records (rows) and/or attributes (columns) of X . The ability to selectively manipulate microdata values within subsets of X (i.e., to apply data masks selectively to submatrices of X) is important for disclosure limitation purposes. This can be accomplished by combining elementary matrix masks that enable *subset selection* along rows and columns, or both, in X with elementary matrix masks as presented previously. This is accomplished in three stages.

At the first stage, apply the ignoring mask $\text{Ign}(Q, R) = AXB$, where A is the $n \times n$ dimensional matrix $A = \sum_{i \in Q} U_{ii}$, and B is the $p \times p$ dimensional matrix $B = \sum_{j \in R} U_{jj}$. A leaves the values in the selected rows Q of X unchanged, and replaces all other values by zeroes; B has similar effect on the columns R . At the second stage, apply the appropriate mask or combination of masks M of Section 3 to $\text{Ign}(Q, R)$ to effect the desired changes, yielding $\tilde{X} = M(\text{Ign}(Q, R))$. As M is designed to change only the selected values, then all ignored values – which $\text{Ign}(Q, R)$ replaced by zero – remain zero after applying M . To preserve the dimensions of \tilde{X} , deletion operations are modified to replace values to be deleted by zero. Finally, restore the ignored original values of X by means of

$$\tilde{X} = M(\text{Ign}(Q, R)) + X - \text{Ign}(Q, R).$$

4.2 Blurring

When the operation M is microaveraging, the formulation of Section 4.1 provides a matrix mask for the data mask *blurring* of Strudler, Oh and Scheuren (1986).

4.3 Data Swapping

Data swapping is a method whereby selected data values are exchanged between selected sets of records, in a manner that ensures that certain one, two and higher-way tabulations remain unchanged (Dalenius and Reiss 1982). Setting $M = \text{Reo}(P)$, where the swapping rule is given by a permutation P of the affected records, Section 4.1 yields a matrix mask for data swapping.

5. CONCLUDING COMMENTS

A formulation based on matrix algebra for representing the principal statistical disclosure limitation methods for microdata has been developed. Computational issues, such as for large files, are not addressed. However, the partitioning methods of Section 4.1 could be used to reduce effective computational size when working with extremely large files.

Matrix masks offer a comprehensive framework in which statistical agencies can develop, evaluate and use reliable microdata disclosure limitation software. Such software could be shared among agencies. Exploration of the uses of matrix masks by U.S. statistical agencies has been encouraged by an expert panel (Federal Committee on Statistical Methodology 1994, p. 82). The potential effect of the widespread use of matrix masks would be to standardize the microdata disclosure limitation methods available for use by agencies, while expanding each agency's options to evaluate and apply these methods.

ACKNOWLEDGEMENTS

The author is indebted to Professor George T. Duncan, Carnegie Mellon University, for introducing the concept of matrix masks and for collaborations leading to an earlier version of this paper, and to Sumitra Mukherjee, Duncan's doctoral student, for his critical reading and for developing some of the alternative formulations presented here. Preliminary research on this topic was supported in part by National Science Foundation Grant SES 91-10512. The views expressed are those of the author and are not intended to represent the policies or practices of the U.S. Environmental Protection Agency.

REFERENCES

- BETHLEHEM, J.G., KELLER, W.J., and PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.
- COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 398, 520-524.
- COX, L. (1991). Comment (on Duncan, G.T. and R.W. Pearson 1991, below), *Statistical Science*, 6, 232-234.
- COX, L., and ERNST, L. (1982). Controlled rounding. *INFOR*, 20, 423-432.
- DALENIUS, T., and REISS, S. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.

- DUNCAN, G.T. (1990). Inferential disclosure-limited microdata dissemination. *Proceedings of the Survey Research Section, American Statistical Association*, 440-445.
- DUNCAN, G.T., and LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7, 207-217.
- DUNCAN, G.T., and PEARSON, R.W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6, 219-239.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on disclosure limitation methodology. Statistical Policy Working Paper 22, Office of Management and Budget, Washington, DC.
- STRUDLER, M., OH, L., and SCHEUREN, F. (1986). Protection of taxpayer confidentiality with respect to the tax model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 375-381.
- TENDICK, P. (1991). Optimal noise addition for preserving confidentiality in multivariate data. *Journal of Statistical Planning and Inference*, 27, 341-353.

Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey

P.D. FALORSI, S. FALORSI and A. RUSSO¹

ABSTRACT

The study was undertaken to evaluate some alternative small areas estimators to produce level estimates for unplanned domains from the Italian Labour Force Sample Survey. In our study, the small areas are the Health Service Areas, which are unplanned sub-regional territorial domains and were not isolated at the time of sample design and thus cut across boundaries of the design strata. We consider the following estimators: post-stratified ratio, synthetic, composite expressed as linear combination of synthetic and of post-stratified ratio, and sample size dependent. For all the estimators considered in this study, the average percent relative biases and the average relative mean square errors were obtained in a Monte Carlo study in which the sample design was simulated using data from the 1981 Italian Census.

KEY WORDS: Small area estimators; Unplanned domains; Bias; Mean Square Error; Simulation study.

1. INTRODUCTION

In Italy, as in many other countries, there is a growing need for current and reliable data on small areas. This information need concerns most sample surveys realised by the Italian National Statistical Institute (ISTAT), especially the Labour Force Survey (LFS), which has been studied to warrant accuracy in regional estimates.

In the past, ISTAT's solution to this problem was to broaden the sample without changing the estimation method (Fabbris *et al.* 1988). In the last few years, however, in order to find a solution to the negative aspects of over-sized samples, research has been launched to identify estimation methods to improve the accuracy of small areas estimates (Falorsi and Russo 1987, 1989, 1990 and 1991).

In our study, the small areas are the Health Service Areas (HSA), which are unplanned sub-regional territorial domains and were not isolated at the time of sample design and thus cut-across the boundaries of the design strata. The sizes of these territorial domains are such that the reliability of regular estimates would have been satisfactory had these domains been designed with separate fixed sample sizes from individual domains.

The study was undertaken to evaluate some of the alternative small areas estimators to produce HSA level estimates from the LFS.

We consider the following estimators: post-stratified ratio, synthetic, composite (expressed as linear combination of the synthetic and of the post-stratified ratio), and sample size dependent.

For all the estimators considered in this study, the average percent relative biases and the average relative mean square errors were obtained in a Monte Carlo study

in which the LFS design was simulated using data from the 1981 Italian Census.

2. BRIEF DESCRIPTION OF THE LFS SAMPLE STRATEGY

2.1 Design

The LFS is based on a two stage sample design stratified for the primary sampling units (PSU). The PSUs are the municipalities, while the secondary sampling units (SSU) are the households. In the framework of each geographical region the PSUs are divided according to the provinces. In each province the PSUs are divided into two main area types: the self-representing area consisting of the larger PSUs, and the non self-representing area consisting of the smaller PSUs.

All PSUs in the self-representing area are sampled, while the selection of PSUs in the non self-representing area is carried out within the strata that have approximately equal measures of size. Two sample PSUs are selected from each stratum without replacement and with probability proportional to size (total number of persons). The SSUs are selected without replacement and with equal probabilities from the selected PSUs independently. All members of each sample household are enumerated.

2.2 Estimator of Total

With reference to the generic geographical region, we introduce the following subscripts: h , for stratum ($h = 1, \dots, H$); i , for primary sampling unit; j , for secondary sampling units; g , for age-sex groups ($g = 1, \dots, G$).

¹ P.D. Falorsi, Senior Researcher, National Statistical Institute, Rome, Italy; S. Falorsi, Researcher, National Statistical Institute, Rome, Italy; Aldo Russo, Associate Professor, University of Molise, Campobasso, Italy.

In the present study we consider the following age classes 14-19, 20-29, 30-59, 60-64, and over 65.

A quantity referring to stratum h , primary sampling unit i , and secondary sampling unit j will be briefly referred to as the quantity in hij ; and a quantity referring to stratum h and primary sampling unit i will be referred to as the quantity in hi .

The following notations are also used: N_h , for number of PSUs in h ; P_h , for total number of persons in h ; n_h , for number of sample PSUs selected in h ; M_{hi} for number of SSUs in hi ; P_{hi} , for total number of persons in hi ; m_{hi} , for number of sample SSUs selected in hi ; P_{ghij} , for number of persons in group g belonging to hij ; P_{hij} , for number of persons in hij .

Further let

$$Y = \sum_{g=1}^G \sum_{h=1}^H \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ghij}$$

be the total of the characteristic y for regional population, where Y_{ghij} denotes total of the characteristic of interest y for the P_{ghij} persons. Actually, the estimate of Y is obtained by a post-stratified estimator. This estimator is given by:

$$\hat{Y} = \sum_{g=1}^G \frac{\hat{Y}_g}{\hat{P}_g} P_g,$$

where

$$\hat{Y}_g = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij}; \hat{P}_g = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij}$$

represent unbiased estimates of

$$Y_g = \sum_{h=1}^H \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ghij}; P_g = \sum_{h=1}^H \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} P_{ghij}.$$

In the above formulas, the symbol K_{hij} , that denotes the basic weight, is expressed by:

$$K_{hij} = \frac{P_h}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}}.$$

3. SMALL AREA ESTIMATORS

With reference to the generic geographical region, we suppose that the population P is divided into D non-overlapping small areas 1, ..., d , ..., D for which estimates are required. Each area is obtained by an aggregation of municipalities. The problem considered is the estimation the total of a y -variable for all units belonging

to the small area d . In practice, the small area d will have a non-null intersection with only a certain number of design strata which we denote as $\tilde{H} = \{h \mid {}_dP_h > 0\}$, where ${}_dP_h$ represents the part of P_h belonging to the small area d .

Denoting by ${}_dN_h$ the number of PSUs belonging to small area d in stratum h , we seek to estimate the small area total

$${}_dY = \sum_{g=1}^G \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{{}_dN_h} \sum_{j=1}^{M_{hi}} Y_{ghij}.$$

The development of a particular estimation method for small areas basically depends on available information. In Italy the accessible information at small area level is very poor. At present the accessible territorial information is total population by sex for each municipality collected through register statistics. In a future context (at end of 1994), the population counts by age-sex group will be available for each municipality. For this reason, in the present study we consider only those small area estimators that utilize, as auxiliary information, the population total by age-sex group.

3.1 Post-stratified Ratio Estimator

A post-stratified ratio estimator (POS) of ${}_dY$ is given by:

$${}_d\hat{Y}_{POS} = \sum_{g=1}^G \frac{{}_d\hat{Y}_g}{{}_d\hat{P}_g} {}_dP_g, \quad (1)$$

where

$${}_d\hat{Y}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij} \delta_{hi},$$

$${}_d\hat{P}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij} \delta_{hi},$$

$${}_dP_g = \sum_{h=1}^{\tilde{H}} {}_dP_{gh} = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{{}_dN_h} \sum_{j=1}^{M_{hi}} P_{ghij},$$

in which ${}_dP_{gh}$ denotes the total population for the age/sex group g in small area d intersected by stratum h , δ_{hi} is a binary variate that equals 1 if the PSU hi belongs to the small area d and equals 0 otherwise. For a better explanation of formula (1), we observe that PSU is a subset of small area and then does not intersect it.

The post-stratified ratio estimator is unbiased except for the effect of ratio estimation bias which is usually negligible. The estimator is defined to be zero when there is no sample within the domain. This estimator is not reliable for small sample sizes.

3.2 Synthetic Estimator

For computing a synthetic estimator, it is assumed that the small area population means for given population sub-groups are approximately equal to the larger area populations means of the same sub-groups. This estimator is obtained by means of a two steps procedure: (i) with respect to an aggregated territorial level, estimates of the investigated features are determined for population sub-groups; (ii) estimates for the aggregated territorial level area are then scaled in proportion to the sub-group incidence within the small domain of interest.

The synthetic estimator has a low variance since it is based on a larger sample, but it suffers from bias depending on the distance from the assumption of homogeneity, for each subgroup, between the small area and the larger area with reference to the characteristic of interest, y . The problems associated with synthetic estimators have been documented by Purcell and Linacre (1976), Gonzalez and Hoza (1978), Ghangurde and Singh (1978), Schaible (1979) and Levy (1979) among others.

In this study we consider the following form of synthetic estimator (SYN):

$${}_d\hat{Y}_{\text{SYN}} = \sum_{g=1}^G \frac{\tilde{Y}_g}{\tilde{P}_g} {}_dP_g, \quad (2)$$

where

$$\tilde{Y}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ghij}; \quad \tilde{P}_g = \sum_{h=1}^{\tilde{H}} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ghij}.$$

3.3 Composite Estimator

The composite estimator (COM) considered here is obtained as a linear combination of the estimators SYN (biased with low sample variance) and POS (less biased with high sample variance):

$${}_d\hat{Y}_{\text{COM}} = \alpha {}_d\hat{Y}_{\text{POS}} + (1 - \alpha) {}_d\hat{Y}_{\text{SYN}}, \quad (3)$$

where α is a constant ($0 \leq \alpha \leq 1$). This estimator minimizes the chances of extreme situations (both in terms of bias and sample variance). Therefore, in a given concrete situation such estimator may turn out to be more advantageous than its two components considered separately.

The optimum value for α that minimizes the MSE of the COM estimator is given by

$$\alpha_{\text{opt}} = \frac{\text{MSE}({}_d\hat{Y}_{\text{SYN}}) - E({}_d\hat{Y}_{\text{SYN}} - {}_dY)({}_d\hat{Y}_{\text{POS}} - Y_d)}{\text{MSE}({}_d\hat{Y}_{\text{SYN}}) + \text{MSE}({}_d\hat{Y}_{\text{POS}}) - 2E({}_d\hat{Y}_{\text{SYN}} - {}_dY)({}_d\hat{Y}_{\text{POS}} - Y_d)}. \quad (4)$$

Furthermore, when neglecting the covariance term in (4), under the assumption that this term will be small relative to $\text{MSE}({}_d\hat{Y}_{\text{SYN}})$ and $\text{MSE}({}_d\hat{Y}_{\text{POS}})$, the optimal weight α can be approximated by

$$\alpha_{\text{opt}}^* = \frac{\text{MSE}({}_d\hat{Y}_{\text{SYN}})}{\text{MSE}({}_d\hat{Y}_{\text{SYN}}) + \text{MSE}({}_d\hat{Y}_{\text{POS}})}. \quad (5)$$

This is the approach to define weights followed by Schaible (1978).

In our work the optimal values of α have been obtained from Census data using formula (5). When considering a real sample survey only an estimated value of optimum α may be used, thus resulting in a decrease in efficiency.

3.4 Sample Size Dependent Estimator

The sample size dependent estimator is a particular case of the composite estimator. The linear combination of synthetic and of the less biased estimator is made for each sub-group and depends on the outcome of the given sample. We consider the following form of sample size dependent estimator (SD) which take into account the realized sample size in the small area. It is defined as (Drew, Singh and Choudhry 1982):

$${}_d\hat{Y}_{\text{SD}} = \sum_{g=1}^G \left\{ \alpha_g \left(\frac{{}_d\hat{Y}_g}{{}_d\hat{P}_g} {}_dP_g \right) + (1 - \alpha_g) \frac{\tilde{Y}_g}{\tilde{P}_g} {}_dP_g \right\}, \quad (6)$$

where

$$\alpha_g = \begin{cases} 1/({}_dR_g F) & 1/{}_dR_g < F, \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

with ${}_dR_g = {}_dP_g / {}_d\hat{P}_g$.

The constant F is chosen to control the contribution of the synthetic component. The reliance on the synthetic portion decreases as the value of F increases. The choice of the value for F would depend upon several factors. In our study the efficiency of sample dependent estimator has been investigated for $F = 1$. This value proved to be efficient while affording protection against the bias of synthetic estimator.

The logic behind the SD estimator is that when the sample size within domain d and group g is small, then the direct estimate for domain d and group g would be unstable and a synthetic estimate may be superior. However, if the sample in domain d and group g is larger than expected this is not a problem, since the performance of the post-stratified direct part would improve as the sample size improves. In conclusion, we observe that SD estimator may be considered as a particular form of sample size dependent regression estimator given in Särndal and Hidiroglou (1989), that has good conditional properties.

4. DESCRIPTION OF THE EMPIRICAL STUDY

4.1 Simulation of the LFS Sample Design

In our study, we have considered the 14 HSAs of the Friuli region as small areas. The variable of interest, y , is the number of unemployed.

Evaluation of the performance of the various estimators, discussed in Section 3, was done by referring to a sample design (two stages with stratification of the PSUs) identical to that adopted for the LFS in Friuli. This design is based on the selection of 39 PSUs and 2,290 SSUs from a population of 219 PSUs and 465,000 SSUs.

We have selected independently 400 Monte Carlo sample replicates each of identical size (in terms of PSUs and of SSUs) of the LFS' sample. All the information utilized in the simulation is taken from the 1981 General Population Census, so ${}_dY$ is known.

4.2 Evaluation of Small Area Estimators

We denote by ${}_d\hat{Y}(mr)$ the estimate of the total ${}_dY$ for the small area d from the r th Monte Carlo replicate when using the estimator m . The percent relative bias of estimator m for the small area d is given by

$${}_d\text{ARB}_m = \frac{1}{R} \left(\sum_{r=1}^R \frac{{}_d\hat{Y}(mr)}{{}_dY} - 1 \right) 100,$$

where R is the number of samples ($R = 400$).

The average of the percent absolute relative bias of estimator m over the whole set of small areas is:

$$\overline{\text{ARB}}_m = \frac{1}{D} \sum_{d=1}^D |{}_d\text{ARB}_m|,$$

where D is the number of small areas under observation ($D = 14$).

The percent root mean square error of estimator m for small area d is

$${}_d\text{RMSE}_m = \frac{\sqrt{{}_d\text{MSE}_m}}{{}_dY} 100,$$

where the mean square error of estimator m for the small area d is expressed by

$${}_d\text{MSE}_m = \frac{1}{R} \sum_{r=1}^R ({}_d\hat{Y}(mr) - {}_dY)^2.$$

The average percent root mean square error of estimator m over all areas is

$$\text{RMSE}_m = \frac{1}{D} \sum_{d=1}^D {}_d\text{RMSE}_m.$$

4.3 Analysis of Results

A. Overall Performance Measures

The average percent absolute biases and the average percent root mean square errors of the small area estimators for the LFS characteristic “number of unemployed persons” are presented in Table 1. Looking at this table, the following conclusions emerge:

- (i) As expected, POS presents the smallest bias. The bias of SYN is larger than the bias of the other estimators. The bias of COM is roughly 30% lower than the bias of SYN estimator. The bias of SD estimator is only slightly lower than that of POS estimator.
- (ii) SYN and COM have the smallest average percent root mean square errors, but these estimators are affected by a very high bias. POS, with low bias, is, conversely, the less efficient estimator. The average percent root mean square error of SD is approximately 30% higher than those of SYN and COM estimators.

Table 1

Average Percent Absolute Relative Bias $\overline{\text{ARB}}$
and Average Percent Root Mean Square Error $\overline{\text{RMSE}}$
for Unemployed by Estimator

Estimator	$\overline{\text{ARB}}$	$\overline{\text{RMSE}}$
POS	1.75	42.08
SYN	8.97	23.80
COM	6.00	23.57
SD	2.39	31.08

B. Performance Measures by Small Area

Tables 2 and 3 present the Percent Relative Bias (${}_d\text{ARB}$) and the Percent Root Mean Square Error (${}_d\text{RMSE}$) of the estimators for each of fourteen Health Service Areas in Friuli. Furthermore, Table 2 gives the percent ratio between the population of the HSA and the population of the set \tilde{H} of strata including the HSA (p_1); Table 3 shows the percent ratio between the population of the HSA and the population of the region Friuli (p_2) and the percent ratio between the population of the set \tilde{H} of strata including the HSA and the population of the region Friuli (p_3). Looking at these Tables, the following conclusions emerge:

- (i) SYN and COM are badly biased in some small areas, namely, in those small areas where the model underlying SYN fits poorly. Generally the small areas with low values of the ratio p_1 are affected by large bias (e.g., HSAs 1, 2, 3, 4 and 6). Conversely, large values of the ratio p_1 are associated with low values of the bias (e.g., HSAs 5, 9, 10 and 13). However, SYN and COM consistently have an attractively low RMSE compared to other alternatives. In three of the fourteen areas (viz, areas 3, 4 and 8) COM is consistently the most efficient estimator. In two areas (10 and 12)

SYN is evidently more efficient and in the remaining areas the two estimators are roughly similar from the point of view of efficiency. Furthermore, we observe that the lowest values of RMSE for SYN generally are associated with the highest values of the ratio p_3 (e.g., HSAs 1, 2, 5, 6, 9 and 13). HSAs 3 and 4, while having an high value of the ratio p_3 , present a high value of RMSE. This is due to the large bias.

- (ii) POS shows negligible bias values in almost all small areas. The RMSE values of POS are much higher than those of the other estimators in all the small areas. We observe that the RMSE of the POS estimator is negatively correlated with the ratio p_2 . This is caused by the fact that the expected sample size increases as the ratio p_2 increases. Consequently, the variance (which is the main component of MSE of POS) decreases.
- (iii) The estimator SD presents a negligible bias in seven (5, 7, 9, 10, 11, 12 and 13) of the fourteen small areas. In the other areas the bias is quite low. Furthermore, in nine areas (2, 3, 4, 5, 9, 10, 11, 12 and 13) SD has a bias similar to that of POS. The estimator SD is better, from the MSE point of view, in comparison with POS. In four areas (7, 8, 9, and 13) RMSE is similar to those of SYN and COM.
- (iv) Finally, we notice that in the largest areas with the highest values of the ratio p_2 (e.g., HSAs 9 and 5) all the estimators considered give similar results in terms of bias and MSE. For the remaining areas, where the estimators have different performances, there is a problem in the choice of the best estimator.

Table 2

Percent Relative Bias ($\%ARB$) of Each of Fourteen Health Service Areas (HSA) in Friuli for Unemployed by Estimator

HSA	p_1	Estimator			
		POS	SYN	COM	SD
1	19.1	-1.57	-10.92	-7.68	-3.01
2	16.1	-5.61	-9.21	-6.97	-4.79
3	15.3	-5.21	28.82	17.98	5.79
4	16.3	-2.50	20.92	15.02	2.99
5	47.1	-0.46	1.61	0.98	-0.28
6	24.6	-1.37	-12.24	-9.06	-3.28
7	81.8	0.05	-6.25	-3.40	-1.66
8	70.7	0.81	11.80	6.63	2.17
9	92.2	0.47	0.76	0.68	0.78
10	71.2	0.36	-1.34	0.51	-1.02
11	21.7	-1.01	-5.64	-5.00	-1.62
12	40.6	-1.52	-6.66	-6.05	-1.19
13	56.3	-0.95	-3.12	-1.11	-1.28
14	21.8	-2.51	-6.21	-3.03	-3.53

p_1 = percent ratio between the population of the HSA and the population of the set \bar{H} of strata including the HSA.

Table 3

Percent Root Mean Square Error ($\%RMSE$) of Each of Fourteen Health Service Areas (HSA) in Friuli for Unemployed by Estimator

HSA	p_2	p_3	Estimator			
			POS	SYN	COM	SD
1	3.8	19.9	52.23	20.41	21.12	32.39
2	3.1	19.2	63.36	19.45	20.81	38.30
3	3.6	23.2	57.44	36.57	30.71	42.46
4	3.8	23.2	58.19	30.09	27.02	36.88
5	20.2	42.9	18.81	13.38	14.01	17.87
6	8.5	34.8	28.09	17.49	17.00	22.69
7	6.9	8.4	23.83	21.47	21.67	22.67
8	4.8	6.8	28.75	28.54	26.35	27.40
9	21.2	22.9	17.29	16.15	16.40	16.89
10	1.8	2.5	67.00	50.12	53.31	59.27
11	3.2	14.6	49.82	18.35	19.20	30.42
12	4.3	10.7	46.40	22.10	24.04	33.18
13	12.6	22.4	20.13	15.53	15.40	17.88
14	2.3	10.1	57.80	23.58	22.94	36.81

p_2 = percent ratio between the population of the HSA and the population of the region Friuli.

p_3 = percent ratio between the population of the set \bar{H} of strata including the HSA and the population of the region Friuli.

5. CONCLUSIONS

From the point of view of bias, the post-stratified ratio estimator (POS) is essentially unbiased in almost all the small areas. Furthermore the sample size dependent estimator (SD) has negligible values of the bias in almost all small areas. Synthetic (SYN) and composite (COM) estimators present bias values much higher than those of the other estimators.

From the point of view of efficiency, SYN and COM consistently have significantly lower RMSE compared to other alternatives. The estimator SD is much more efficient than POS and furthermore in four of the fourteen areas it shows RMSE values close to those of SYN and COM. Further, when considering the estimator COM there is the problem of the computation of optimum α . In practice only an estimated value of α may be used, resulting in a decrease in efficiency of this estimator. Thus considering both, bias and efficiency, the SD estimator would seem to be preferable to other estimators examined in the context of LFS in Friuli. The sampling rates in Friuli are relatively high and the magnitudes of relative biases and efficiencies of these estimators may be different in other regions where the sampling rates are low, e.g., Piemonte and Lombardia.

REFERENCES

- DREW, J.D., SINGH, M.P., and CHOUDHRY, G.H. (1982). Evaluation of small area techniques for the Canadian Labour Force Survey. *Survey Methodology*, 8, 17-47.
- FABBRIS, L., RUSSO, A., and SANETTI, I. (1988). Storia e proposte in tema di campionamento a livello regionale, provinciale e sub-provinciale per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 4. Dipartimento di Scienze Statistiche, Università di Padova.
- FALORSI, P.D., and RUSSO, A. (1987). Un metodo di stima sintetica per piccoli domini territoriali nelle indagini ISTAT sulle famiglie. *Atti del Convegno della Società Italiana di Statistica*, Perugia, Italia, 11-20.
- FALORSI, P.D., and RUSSO, A. (1989). Un'analisi comparativa di alcune tecniche di stima per piccole aree per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 18. Dipartimento di Scienze Statistiche, Università di Padova.
- FALORSI, P.D., and RUSSO, A. (1990). La stima dell'errore quadratico medio di alcune forme di stimatore sintetico nei campioni a due stadi utilizzati nelle indagini ISTAT sulle famiglie. *Giomate di studio: Classificazione ed analisi dei dati, metodi, software, applicazioni*, Pescara, Italia, 27-39.
- FALORSI, P.D., and RUSSO, A. (1991). Evaluation of small area estimation techniques for Italian Labour Force Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 80-106.
- GHANGURDE, P.D., and SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 52-61.
- GONZALEZ, M.E., and HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.
- LEVY, P.S. (1979). Small area estimation synthetic and other procedures, 1968-1978. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 4-19.
- PURCELL, N.J., and LINACRE, S. (1976). Techniques for the Estimation of Small Area Characteristics. Paper presented at the 3rd Australian Statistical Conference, Melbourne.
- SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.
- SCHAIBLE, W.L. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.
- SCHAIBLE, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 36-83.

Nonparametric Estimation of Response Probabilities in Sampling Theory

THÉOPHILE NIYONSENGA¹

ABSTRACT

We deal with the nonresponse problem by drawing on the model of selection in phases that was proposed by Särndal and Swenson (1987). To estimate response probabilities, we use the nonparametric approach first advanced by Giommi (1987). We define estimators according to the nonparametric estimation (NPE) model, and we study their general properties empirically. Inference is based on the concept of quasi-randomization (Oh and Scheuren 1983). The emphasis is on estimating the variance and constructing confidence intervals. We find, by way of a Monte Carlo study, that it is possible to improve the quality of the estimators considered by using a variant of the NPE approach. The latter also serves to confirm the performance of regression estimators in terms of variance estimation.

KEY WORDS: Weighting by phases; Regression estimator; Variance estimators.

1. INTRODUCTION

To counter the effect of nonresponse on the estimation of parameters of a finite population, we consider the phenomenon of nonresponse as a unit selection process in three phases. We therefore use weighting by phases. This adjustment procedure assigns to each unit observed a weight that is inversely proportional to the probability of appearing in the sample, to the unit response probability given the sample, and to the item response probability given the sample and the set of respondents per unit.

In practice, only the probabilities of inclusion in the sample are known. The problem facing us is to estimate individual response probabilities before incorporating them in formulas for the estimators of interest. The nonparametric estimation approach is one of the response probability estimation procedures. It is motivated by the use of auxiliary variables which are linked with unit and item response mechanisms (Giommi 1985, 1987), and which may be correlated with the variables of interest. This avoids assuming that nonresponse is independent of the variables being studied (Oh and Scheuren 1983). This approach also enables us to avoid postulating one or more parametric models governing response, such as the Logit and Tobit models (Grosbras 1987b; Chicoineau, Payen and Thélot 1985) or models of uniform response within subpopulations (Oh and Scheuren 1983; Särndal and Swenson 1985, 1987).

In the Monte Carlo study illustrating certain estimators according to the nonparametric approach, we consider the quite specific case in which the two response mechanisms are governed by the same auxiliary variables. The difference between items will reside in the degree of correlation between each item and the auxiliary variables.

2. NONRESPONSE: A THREE-PHASE SELECTION PROCESS

Consider a finite population $U = \{1, 2, \dots, k, \dots, N\}$, of size N . Let s be a sample of fixed size n drawn from U according to a plan $\mathcal{P}(s)$ known and characterized by inclusion probabilities $\pi_k > 0, \forall k$ and $\pi_{k\ell} > 0 \forall k \neq \ell$. We want to observe the units $k \in s$ in relation to a set of Q items $y_1, \dots, y_q, \dots, y_Q$ ($Q \geq 1$), then estimate the total per item $t_q = \sum_U y_{qk}$, for every q ($q = 1, \dots, Q$). We assume that conditional on s , each unit k has a probability $\varphi_k > 0$ of participating in the survey and that the probability that two units k and ℓ participate is $\varphi_{k\ell} > 0$ with $\varphi_{kk} = \varphi_k$. We denote the set of units that agree to participate in the survey by r and the mechanism by which the set r was obtained by $\mathcal{P}(r | s)$. We further assume that conditional on s and r , each unit $k \in r$ responds to item y_q with probability $\psi_{qk} > 0$ and that the probability that two units k and $\ell \in r$ respond to item y_q is $\psi_{qk\ell} > 0$ with $\psi_{qkk} = \psi_{qk}$. We denote by r_q the set of units that, having agreed to participate in the survey, respond to item y_q and by $\mathcal{P}(r_q | s, r)$ the mechanism by which the set r_q is obtained for all q ($q = 1, \dots, Q$).

The sets s , r and r_q are obtained from three selection phases for which only the probabilities of inclusion in s are known. The composition of the unit selection mechanisms gives rise to probability outputs that we denote by $\pi_k \Theta_{qk}$ where $\Theta_{qk} = \varphi_k \psi_{qk}$ and $\Theta_{qk\ell} = \varphi_{k\ell} \psi_{qk\ell}$ with $\Theta_{qkk} = \Theta_{qk}$, which do not correspond to inclusion probabilities. Nor does the quantity Θ_{qk} correspond to an inclusion probability for the two response phases conditional on s . If we define the probabilities of inclusion in r_q by $\pi_{qk}^* = \mathbb{P}(k \in r_q)$ and the probabilities of inclusion in r_q given s by $\Theta_{qk}^* = \mathbb{P}(k \in r_q | s)$, then (i) $\pi_{qk}^* \neq \pi_k \Theta_{qk}^*$

¹ Théophile Niyonsenga, Ph.D., Researcher, Centre de Recherche Clinique, Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, QC, Canada, J1H 5N4.

and (ii) $\pi_k \Theta_{qk}^* \neq \pi_k \Theta_{qk}$. Furthermore, (iii) $\Theta_{qk}^* = \Theta_{qk}$ if probabilities ψ_{qk} are independent of r , and (iv) $\pi_{qk}^* = \pi_k \Theta_{qk}$ if the φ_k do not depend on s and if the ψ_{qk} do not depend on either r or s .

3. A FEW SPECIAL ESTIMATORS

Assume that there is an auxiliary variable x_q (for the q -th item) strongly correlated with the variable y_q and such that x_{qk} is known $\forall k \in s$ or $\forall k \in U$. We take the specific case in which $x_{qk} = x_k$, $\forall q (q = 1, \dots, Q)$, and we assume the following linear model ξ

$$\begin{cases} \mathbb{E}_\xi(y_{qk} | x_k) = \beta_q x_k \\ \text{Cov}_\xi(y_{qk}, y_{q\ell} | x_k, x_\ell) = \begin{cases} \sigma_q^2 x_k & \text{if } k = \ell, \\ 0 & \text{otherwise} \end{cases} \end{cases} \quad (3.1)$$

in which β_q and σ_q are unknown parameters. The following results are extensions of the findings of Särndal and Swenson (1987).

Result 1. If x_k is known, $\forall k \in s$, then the regression estimator, denoted by \hat{t}_{Reg} and defined by:

$$\hat{t}_{\text{Reg}} = \left(\sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \right) \sum_s \frac{x_k}{\pi_k}, \quad (3.2)$$

is approximately unbiased for t_q . Its approximate variance is a sum of three components V_1 , V_2 and V_3 representing the respective portions of the variance due to the selection phases, that is:

$$V_1 = \sum \sum_U \Delta_{\pi_{k\ell}} (y_{qk}/\pi_k) (y_{q\ell}/\pi_\ell),$$

$$V_2 = \mathbb{E} \left\{ \sum \sum_s \Delta_{\varphi_{k\ell}} (E_{qk}/\pi_k \varphi_k) (E_{q\ell}/\pi_\ell \varphi_\ell) \right\},$$

$$V_3 = \mathbb{E} \mathbb{E} \left[\sum \sum_r \Delta_{\psi_{qk\ell}} (E_{qk}/\pi_k \Theta_{qk}) (E_{q\ell}/\pi_\ell \Theta_{q\ell}) | s \right],$$

where the E_{qk} are theoretical residuals of model (3.1). An estimator of $V(\hat{t}_{\text{Reg}})$ is given by $\hat{V}(\hat{t}_{\text{Reg}}) = \hat{V}_1 + \hat{V}_2^+$ (where $\hat{V}_2^+ = \hat{V}_2 + \hat{V}_3$) with:

$$\hat{V}_1 = \sum \sum_{r_q} \frac{\Delta_{\pi_{k\ell}}}{\pi_{k\ell} \Theta_{qk\ell}} \left(\frac{y_{qk}}{\pi_k} \right) \left(\frac{y_{q\ell}}{\pi_\ell} \right), \quad (3.3)$$

and

$$\hat{V}_2^+ = \sum \sum_{r_q} \frac{\Delta_{\varphi_{k\ell}}}{\Theta_{qk\ell}} \left(\frac{e_{qk}}{\pi_k \Theta_{qk}} \right) \left(\frac{e_{q\ell}}{\pi_\ell \Theta_{q\ell}} \right), \quad (3.4)$$

where $\Delta_{\pi_{k\ell}} = \pi_{k\ell} - \pi_k \pi_\ell$, $\Delta_{\varphi_{k\ell}} = \varphi_{k\ell} - \varphi_k \pi_\ell$, $\Delta_{\psi_{qk\ell}} = \psi_{qk\ell} - \psi_{qk} \psi_{q\ell}$ and $\Delta_{\Theta_{qk\ell}} = \Theta_{qk\ell} - \Theta_{qk} \Theta_{q\ell}$, the e_{qk} being the observed residuals obtained from model (3.1).

Result 2. If x_k is known, $\forall k \in U$, then the regression estimator, denoted by \hat{t}_{Reg1} and defined by:

$$\hat{t}_{\text{Reg1}} = N \bar{x}_U \left(\sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \right) \left(\sum_{r_q} \frac{x_k}{\pi_k \Theta_{qk}} \right), \quad (3.5)$$

is approximately unbiased for t_q . Its approximate variance is also a sum of three components V_1 , V_2 and V_3 . The expression of $V_1(\hat{t}_{\text{Reg1}})$ differs from that of $V_1(\hat{t}_{\text{Reg}})$ by the use of the theoretical residuals E_{qk} in place of the raw values y_{qk} , whereas the expressions of V_2 and V_3 are identical to those defined above for \hat{t}_{Reg} . An estimator of $V(\hat{t}_{\text{Reg1}})$ is given by $\hat{V}(\hat{t}_{\text{Reg1}}) = \hat{V}_1 + \hat{V}_2^+$ where:

$$\hat{V}_1 = \sum \sum_{r_q} \frac{\Delta_{\pi_{k\ell}}}{\pi_{k\ell} \Theta_{qk\ell}} \left(\frac{e_{qk}}{\pi_k} \right) \left(\frac{e_{q\ell}}{\pi_\ell} \right), \quad (3.6)$$

and where $\hat{V}_2^+ = \hat{V}_2 + \hat{V}_3$ is obtained by the formula (3.4).

Comment 1. If $x_k = 1$, $\forall k \in U$, the formula (3.5) defines an estimator, denoted by \hat{t}_{Exp} where:

$$\hat{t}_{\text{Exp}} = N \sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}} \left/ \sum_{r_q} \frac{1}{\pi_k \Theta_{qk}} \right. = \frac{N}{\tilde{N}} \sum_{r_q} \frac{y_{qk}}{\pi_k \Theta_{qk}}. \quad (3.7)$$

The estimator \hat{t}_{Exp} is called an “expansion estimator”. An estimator of approximately unbiased variance for $V(\hat{t}_{\text{Exp}})$ is derived from formulas (3.4) and (3.6).

Comment 2. If we take $\Theta_{qk} = \Theta_q (0 < \Theta_q \leq 1)$, $\forall k \in U$, in formula (3.7), we obtain an estimator, denoted by \hat{t}_{Naive} , called a “naive estimator”. Its expression is given by:

$$\hat{t}_{\text{Naive}} = N \sum_{r_q} \frac{y_{qk}}{\pi_k} \left/ \sum_{r_q} \frac{1}{\pi_k} \right. \quad (3.8)$$

If the π_k are constant, the expression (3.8) becomes identical to formula (3.5) in which t is assumed that $\Theta_{qk} = \Theta_q (0 < \Theta_q \leq 1)$, $\forall k \in U$, and $x_k = 1$, $\forall k \in U$.

Comment 3. For the four estimators defined above, the underlying models are derived from model (3.1) and are the following: $y_{qk} = \beta_q x_k + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ and $V(\epsilon_{qk}) = \sigma_q^2 x_k$ for the first two, $y_{qk} = \beta_q + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ and $V(\epsilon_{qk}) = \sigma_q^2$ and N is known for the last two. For the naive estimator, it is necessary to add the uniform unit and item response model.

4. ESTIMATORS WITH ESTIMATED RESPONSE PROBABILITIES

In practice, the response probabilities φ_k and ψ_{qk} as well as the probability outputs $\Theta_{qk} = \varphi_k \psi_{qk}$ ($k \in U$, $q = 1, \dots, Q$) are actually parameters to be estimated. We estimate them by $\hat{\varphi}_k$, $\hat{\psi}_{qk}$ and $\hat{\Theta}_{qk} = \hat{\varphi}_k \hat{\psi}_{qk}$ respectively. We define estimators having the same form as the prototype estimators \hat{t}_{Exp} , \hat{t}_{Reg} and \hat{t}_{Reg1} seen in section 3, taking care to replace the unknown parameters by their respective estimates. We denote these estimators by \hat{t}_{Exp}^* , \hat{t}_{Reg}^* and \hat{t}_{Reg1}^* respectively. The variance estimators are obtained from the expressions (3.3), (3.4) and (3.6), in which the unknown parameters are replaced with their estimates.

4.1 Estimation of Response Probabilities

In theory, the probabilities φ_k and ψ_{qk} are functions of the auxiliary variables, that is, functions of the form $\varphi_k = f_1(v, z_k)$ and $\psi_{qk} = f_2(\mu_q, x_{qk})$ in which the quantities v and μ_q ($q = 1, \dots, Q$) are unknown parameters and where the pair of vectors (z, x_q) , that is, $[(z_1, x_{q1}), \dots, (z_k, x_{qk}), \dots, (z_N, x_{qN})]'$, contain the auxiliary information available for each item y_q . The nonparametric estimation approach uses only the information contained in (z, x_q) to estimate the φ_k and ψ_{qk} . We are considering here the specific case in which the $z_k = x_{qk} = x_k$, $\forall q$ ($q = 1, \dots, Q$), and $\forall k \in s$.

Let $x_s = \{x_k : k \in s\}$, all the auxiliary information relating to the sample. We specify $\tau_s = \{\tau_k : k \in s\}$, a set of functions such that $\tau_k : \mathbb{R}^n \rightarrow \mathbb{R}^1$, for all k in s . We denote by $g_k = \tau_k(x_s)$, $\forall k \in s$, the value of the k -th function evaluated in x_s . We subdivide s in n groups s_k not necessarily disjoint, the respective sizes of which are given by:

$$n_k = \sum_{j \in s} D(g_k - g_j), \quad (k \in s),$$

$$D(g_k - g_j) = \begin{cases} 1 & \text{if } |g_k - g_j| \leq h_k, \\ 0 & \text{otherwise,} \end{cases}$$

for a given constant h_k which may depend on all the values g_k ($k \in s$). The set $s_k = \{j : g_j \in [g_k \pm h_k]\}$, $\forall k \in s$, contains j units, whose values g_j vary little from one to another. This group is called the group whose unit k is the kernel, or simply the k -th group. In other words, s_k is a subset of s for which the values of x fall within the vicinity of $x = x_k$ in the sense of the Euclidian distance that specifies $d(k, j) = |\tau_k(x_s) - \tau_j(x_s)| \leq h_k = h(g_k)$, meaning that $s_k = \{j : d(k, j) \leq h_k\}$. Let $r_k = s_k \cap r$ and $r_{qk} = s_k \cap r_q$. The respective absolute frequencies of these sets are m_k and m_{qk} where:

$$m_k = \sum_{j \in r} D(g_k - g_j), \quad (k \in r);$$

$$m_{qk} = \sum_{j \in r_q} D(g_k - g_j), \quad (k \in r_q, q = 1, \dots, Q).$$

Comment 4. In the general case in which nonresponse is governed by the pair of vectors (z, x_q) with $z \neq x_q$, the τ_k functions would be defined in terms of z in order to estimate the unit response probabilities φ_k and in terms of x_q to estimate the item response probabilities ψ_{qk} . Note that this kernel approach can be generalized to more than one auxiliary variable governing response. For two variables x_1 and x_2 governing nonresponse, we would specify the set $s_k = \{(j_1, j_2) : g_{j_1} \in [g_{k_1} \pm h_{k_1}] \text{ and } g_{j_2} \in [g_{k_2} \pm h_{k_2}]\}$.

Response probabilities φ_k and ψ_{qk} are estimated respectively by the rates:

$$\hat{\varphi}_k = \frac{m_k}{n_k}, \quad \forall k \in r; \quad \hat{\psi}_{qk} = \frac{m_{qk}}{m_k}, \quad \forall k \in r_q, \quad (4.1)$$

whereas the output $\Theta_{qk} = \varphi_k \psi_{qk}$ is estimated by the rate:

$$\hat{\Theta}_{qk} = \hat{\varphi}_k \hat{\psi}_{qk} = m_{qk}/n_k, \quad (k \in r_q, q = 1, \dots, Q), \quad (4.2)$$

which is nothing other than the response rate in the k -th group. This simplification of the estimated output $\hat{\Theta}_{qk} = \hat{\varphi}_k \hat{\psi}_{qk}$ is, however, possible only when the two response mechanisms are governed by the same auxiliary variables.

Two approaches are considered here: the one based on the values of the variable x (npv) and the one based on the ranks of the values of the variable x (npr). The NPE (npv), proposed by Giommi (1987), is obtained by taking $g_k = \tau_k(x_s) = x_k$ ($k \in s$). To offset the possible effect of excessively large and excessively small values of x_s , we introduce a variant that consists in using the ranks of x_s , that is, NPE(npr). We consider the function u such that $u(z) = 1$ if $z \geq 0$ and $u(z) = 0$ if $z < 0$. For any unit k in s , let $u_k = \sum_s u(x_k - x_j)$ = the number of components of x_s that are less than or equal to x_k = the rank of x_k in s . The NPE(npr) is then equivalent to letting $g_k = \tau_k(x_s) = u_k$ ($k \in s$).

4.2 Selection of Interval Limits

The main problem in the NPE approach is the optimum choice of the h_k constants that determine the limits of the intervals $[g_k - h_k; g_k + h_k]$, $\forall k \in s$, that is, a choice of $h_k = h_k(g_s)$ that reduces the bias and mean square error of any estimator using the estimated outputs $\hat{\Theta}_{qk}$ specified in formula (4.2).

According to Giommi (1985, 1987), the terms n_k , m_k and m_{qk} that are used to estimate the response probabilities are, apart from the standardization factors, estimators by the kernel method of the density function according to the

approach of Rosenblatt (1956) for the various series of values of g . As an example, it is easy to demonstrate that:

$$n_k = \sum_{j \in s} D(g_k - g_j) = 2nh(n)\hat{f}_n(g_k),$$

where $h(n) = h(g_k, k \in s)$ is a positive constant that converges toward zero at a quite appropriate rate. The theoretical optimum constant, according to the least mean square error criterion, is given by $h(n) = K_f n^{-1/5}$ where K_f , such as defined by Rosenblatt (1956) and Wegman (1972a and b), is obtained by the expression $K_f = [9f(x)/2 |f''(x)|^2]^{1/5}$.

In practice, $h(n)$ can be obtained only by simulation, since it depends on the density function to be estimated. Gionmi (1985) used $h(n) = 2EI_s n^{-1/3}$ where EI_s is the interquartile range in the sample. Kraft, Lepage and van Eeden (1983) chose $h(n) = C(n)EI_s$ where $C(n) = (K_f/EI_s)n^{-1/5}$. As our choice, we shall adopt $h(n) = C(n)S_{gs}$, where $C(n) = (K_f/S_{gs})n^{-1/5}$ and where S_{gs} is the corrected standard deviation of the values $g_k (k \in s)$. Basing ourselves on the study of Kraft, Lepage and van Eeden (1983), we will empirically determine a value \hat{C}_n of C that is optimal according to the criterion of least bias and least mean square error of the estimator \hat{t}_{Expnp}^* and compare the two versions of the NPE approach.

4.3 Expansion and Regression Estimators

Calculation of the approximate bias and variance of the estimators \hat{t}_{Exp} , \hat{t}_{Reg} and \hat{t}_{Reg1} is simplified by the fact that the probabilities φ_k and ψ_{qk} are assumed to be known. For estimators \hat{t}_{Expnp}^* , \hat{t}_{Regnp}^* and \hat{t}_{Reg1np}^* , these probabilities are estimated by $\hat{\varphi}_k$ and $\hat{\psi}_{qk}$. These probability estimators do not respond to any probability model that would enable us to calculate the bias and the variance conditional on this model. In other words, the sets r_q are generated by unknown response mechanisms for which we estimate the response probabilities by an approach that does not allow for inference conditional on any model underlying the estimation of probabilities.

We would be tempted to resort to Taylor's serial development of the function $1/\hat{\theta}_{qk}$ to justify the approximation of $1/\hat{\theta}_{qk}$ by $1/\theta_{qk}$. In this case, the bias and the variance of \hat{t}_{Expnp}^* , \hat{t}_{Regnp}^* and \hat{t}_{Reg1np}^* would be approached by the approximate bias and variance of \hat{t}_{Expnp} , \hat{t}_{Regnp} and \hat{t}_{Reg1np} . However, for sample sizes that are not sufficiently large, we are in danger of having $1/\hat{\theta}_{qk} \neq 1/\theta_{qk}$ for the majority of the $k \in r_q$, and consequently:

$$V(\hat{t}_{Expnp}^*) \neq V(\hat{t}_{Exp}), V(\hat{t}_{Regnp}^*) \neq V(\hat{t}_{Reg}), \text{ and } V(\hat{t}_{Reg1np}^*) \neq V(\hat{t}_{Reg1}).$$

However, to construct confidence intervals based on \hat{t}_{Expnp}^* , \hat{t}_{Regnp}^* and \hat{t}_{Reg1np}^* , it is necessary to define estimators for their respective variances. Not having explicit

expressions for these variances, it is difficult to define variance estimators and study their properties analytically. The choice of a given estimator is quite difficult to justify. The most natural way of obtaining variance estimators for the variances of \hat{t}_{Expnp}^* , \hat{t}_{Regnp}^* and \hat{t}_{Reg1np}^* is to do a simple substitution of $\theta_{qk} (= \varphi_k \psi_{qk})$, by $\hat{\theta}_{qk} (= \hat{\varphi}_k \hat{\psi}_{qk})$, $\forall k \in r_q$, and of $\theta_{qk\ell}$ by $\hat{\theta}_{qk\ell}$, $\forall k \neq \ell \in r_q$ ($\hat{\theta}_{qk\ell} = \hat{\varphi}_{k\ell} \hat{\psi}_{qk\ell}$), in all the formulas for variance estimators specified for the respective variance estimators of estimators \hat{t}_{Expnp} , \hat{t}_{Regnp} and \hat{t}_{Reg1np} .

5. MONTE CARLO STUDY: COMPARISON OF ESTIMATORS

For simulation purposes, we assume that Bernoulli trials govern each of the response mechanisms (total or partial) and that a simple random sampling without replacement is the sample design used. We consider a vector $(y_1, y_2, y_3)'$ of three items ($Q = 3$) and a variable x containing the auxiliary information. We first generate the $x_k (k \in U)$ by a gamma distribution with parameters a_1 and a_2 . The generation of items y_1, y_2, y_3 is based on the linear model (3.1) and the gamma distribution. More specifically, we generate the $y_{qk} (k \in U \text{ and } q = 1, 2, 3)$ according to a gamma distribution with parameters $a_{1q}(x_k)$ and $a_{2q}(x_k)$ defined by:

$$a_{1q}(x_k) = \frac{\beta_q^2 x_k}{\sigma_q^2}, \quad a_{2q}(x_k) = \frac{\sigma_q^2}{\beta_q},$$

$$\sigma_q^2 = \beta_q^2 a_2 \left\{ \frac{1}{\rho_{xyq}^2} - 1 \right\}, \quad q = 1, 2, 3.$$

The choice of the gamma distribution is based on its general form, which gives rise to a great variety of distributions, and on the fact that it can represent the distribution of various types of populations (Johnson and Kotz 1970, p. 172). We establish *a priori* the parameters a_1 , a_2 , β_q and ρ_{xyq} ($q = 1, 2, 3$), namely:

$$a_1 = 2, \quad a_2 = 10, \quad (\beta_1 \beta_2 \beta_3)' = (0.75 \ 0.65 \ 0.60)',$$

$$(\rho_{xy1} \rho_{xy2} \rho_{xy3})' = (0.90 \ 0.85 \ 0.70)'.$$

To generate the unit and item response probabilities, we consider the following exponential forms:

$$\varphi_k = \exp\{-(\lambda_1 x_k + \lambda_2 v_k)\} \quad \text{and}$$

$$\psi_{qk} = \exp\{-(\lambda_{1q} x_k + \lambda_{2q} v_{qk})\},$$

where the v_k and the v_{qk} result from a uniform distribution $(0; 1)$. The constants λ_1 , λ_2 , λ_{1q} and λ_{2q} are such that: $\lambda_1 = 0.15/\bar{x}_U$, $\lambda_{1q} = 0.15/\beta_q \bar{x}_U$ and $\lambda_2 = \lambda_{2q} = 0.45$ ($q = 1, 2, 3$). Such a parameterization makes it possible to have an average response rate (total or partial) of approximately 70%. We could have varied these constants or used other continuous functions.

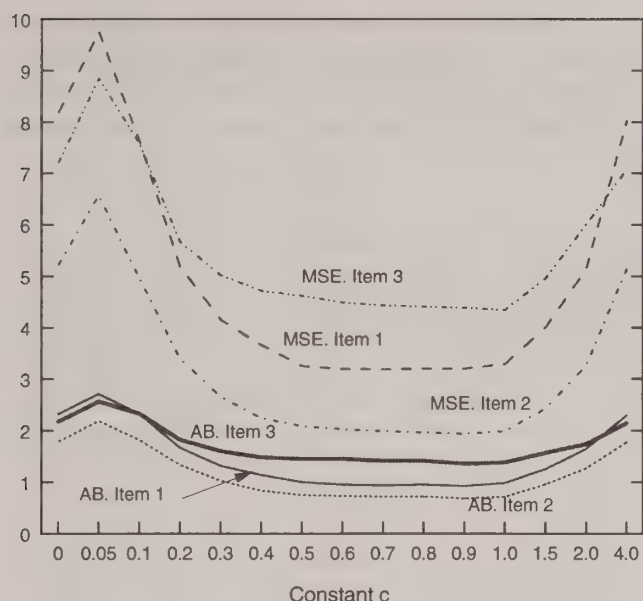


Figure 5.1 Absolute bias and MSE: the estimator \hat{t}_{Expnp}^* for $n = 60$

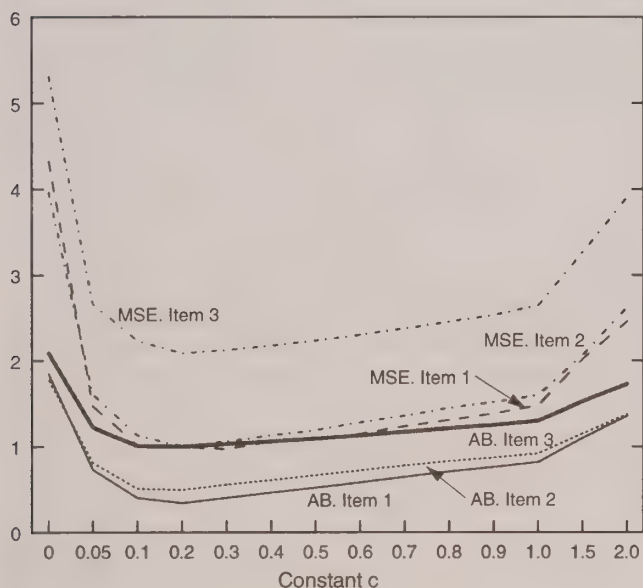


Figure 5.2 Absolute bias and MSE: the estimator \hat{t}_{Expnpr}^* for $n = 200$

5.1 Comparison of the Two Variants of the NPE Approach

We consider a population of size $N = 100$ and draw a sample s of size $n = 60$, which we subject to the response mechanisms. We repeat the sampling IK times and calculate the bias $IB(\hat{t}_{Expnp}^*)$ and the mean-square error $MSE(\hat{t}_{Expnp}^*)$, for different values of C ($C \geq 0$). Next we repeat this experiment with $N = 1,000$ and $n = 200$.

The results of this empirical study are illustrated by the diagrams of $IB(\hat{t}_{Expnp}^*)$ et $MSE(\hat{t}_{Expnp}^*)$ as a function of the constant C . From this brief study we observe, firstly, that the value \hat{C}_n of the optimal constant C is in the interval $[0; 1]$, depends on the size of the sample and decreases as the sample size increases (Figures 5.1 and 5.2).

We also observe that the estimator \hat{t}_{Expnpr}^* is still better in terms of less bias and mean square error than the estimator \hat{t}_{Expnp}^* in the interval $[0; 1]$ as illustrated as an example in Figure 5.3 for item 3, the item the least correlated with the auxiliary variable. A very important fact to be noted is that for the estimator \hat{t}_{Expnpr}^* we more quickly reach the values of the bias and the mean square error of the estimator \hat{t}_{Naive} in $[0; 1]$ at $C = 0.05$ and outside this interval at $C = 4$. Unlike with the estimator \hat{t}_{Expnp}^* , the values of the bias and the mean square error of the estimator \hat{t}_{Expnp}^* first reach maximum values at $C = 0.05$ before taking on the values of the bias and mean-square error of \hat{t}_{Naive} at $C = 0$. We also note that for a fairly large size n and for any value of C in the interval $[0; 1]$, the variation is hardly perceptible (Figure 5.3). For this reason, we suggest that a compromise value be used: $C = 0.5$ (that is, $h = 0.5S_{gs}$).

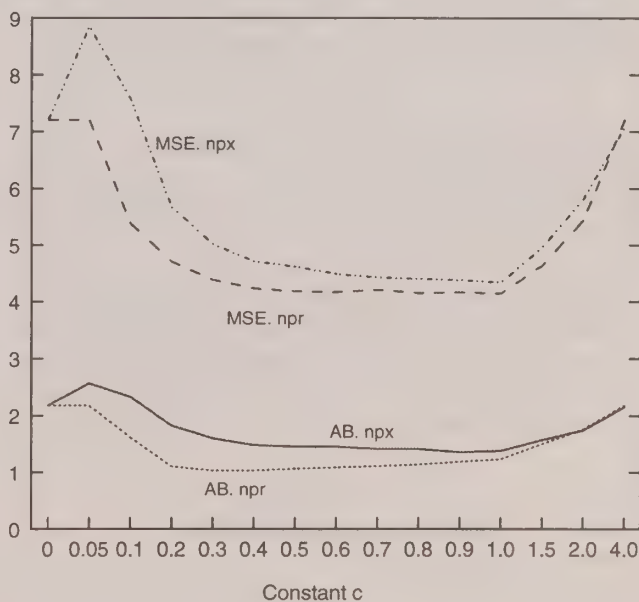


Figure 5.3 Absolute bias and MSE: the estimators \hat{t}_{Expnp}^* and \hat{t}_{Expnpr}^* for item 3

5.2 Overall Comparison of Estimators

The complete operation of the simulation consists in (i) first, drawing the sample s of size $n = 200$ of the population of size $N = 1,000$, (ii) then applying the unit and item response mechanisms to obtain sets r_q ($q = 1, 2, 3$), and (iii) lastly, calculating, for each estimator, the values

of \hat{t} and $\hat{V}(\hat{t})$. We repeat this operation \mathbb{K} times. Once the experiment is completed, we calculate, as performance measurements, (i) the bias $\text{IB}(\hat{t}) = \mathbb{E}(\hat{t}) - t_q$, (ii) the mean square error $\text{MSE}(\hat{t}) = \mathbb{E}(\hat{t} - t_q)^2$, (iii) the expectation of the variance estimator $\mathbb{E}(\hat{V}(\hat{t}))$ and (iv) the theoretical recovery rate $P_o(\hat{t}) = \mathbb{P}\{|\hat{t} - t_q| \leq Z_{\alpha/2} [V(\hat{t})]^{1/2}\}$. We can also calculate, for each given estimator, (v) the relative error $\text{RE}(\hat{t}) [= \text{IB}(\hat{t})/t]$, (vi) the variance $V(\hat{t}) [= \text{MSE}(\hat{t}) - (\text{IB}(\hat{t}))^2]$, (vii) the relative bias $\text{RB}(\hat{t}) [= |\text{IB}(\hat{t})| / (V(\hat{t}))^{1/2}]$ as well as (viii) the relative error of the variance estimator $\text{RE}(\hat{V}(\hat{t})) [= \text{IB}(\hat{V}(\hat{t}))/V(\hat{t})]$ in order to examine the sensitivity of the variance estimators to nonresponse.

5.3 Interpretation of the Results of the Global Simulation

I. The Prototype Estimators

The simulation results confirm the theory. For these estimators, we make the following observations, based on Tables 5.1 to 5.4:

- (i) \hat{t}_{Exp} , \hat{t}_{Reg} and \hat{t}_{Reg1} are approximately unbiased;
- (ii) $\text{MSE}(\hat{t}_{\text{Reg1}}) < \text{MSE}(\hat{t}_{\text{Reg}}) < \text{MSE}(\hat{t}_{\text{Exp}})$;
- (iii) $V(\hat{t}_{\text{Reg1}}) < V(\hat{t}_{\text{Reg}}) < V(\hat{t}_{\text{Exp}})$ and $\mathbb{E}[\hat{V}(\hat{t}_{\text{Reg1}})] < \mathbb{E}[\hat{V}(\hat{t}_{\text{Reg}})] < \mathbb{E}[\hat{V}(\hat{t}_{\text{Exp}})]$.

For these estimators, we also expected that:

- (i) $\mathbb{E}\hat{V}(\hat{t}_{\text{Exp}}) \approx V(\hat{t}_{\text{Exp}})$, $\mathbb{E}\hat{V}(\hat{t}_{\text{Reg}}) \approx V(\hat{t}_{\text{Reg}})$ and $\mathbb{E}\hat{V}(\hat{t}_{\text{Reg1}}) \approx V(\hat{t}_{\text{Reg1}})$;
- (ii) Negligible relative bias [$\text{RB}(\hat{t}) < 0.10$]; the recovery rates are close to the theoretical rates. The relative errors $\text{RE}(\hat{t})$ and $\text{RE}(\hat{V}(\hat{t}))$ are negligible, and are in part due to the simulation (errors due to the limited number of repetitions of the experiment).

Table 5.1
The Values of $\text{IB}(\hat{t})$, $\text{MSE}(\hat{t})$

	y_1		y_2		y_3	
\hat{t}_{Exp}	-0.036	1.690	-0.052	1.525	-0.056	2.299
\hat{t}_{Reg}	-0.020	0.735	-0.019	0.744	-0.030	1.446
\hat{t}_{Reg1}	-0.012	0.319	-0.012	0.431	-0.021	1.202
\hat{t}_{Naive}	-2.037	5.069	-1.937	4.535	-2.220	5.911
\hat{t}_{Expnp}^*	-0.690	1.345	-0.777	1.407	-1.228	2.604
$\hat{t}_{\text{Expnpr}}^*$	-0.601	1.175	-0.709	1.249	-1.140	2.345
$\hat{t}_{\text{Regnpr}}^*$	-0.293	0.785	-0.414	0.830	-0.895	1.834
$\hat{t}_{\text{Reg1npr}}^*$	-0.285	0.376	-0.407	0.520	-0.886	1.621

Table 5.2

The Values of $V(\hat{t})$, $\mathbb{E}[\hat{V}(\hat{t})]$ and $100 \cdot \mathbb{E}[\hat{V}_1(\hat{t})]/\mathbb{E}[\hat{V}(\hat{t})]$

	y_1			y_2			y_3		
\hat{t}_{Exp}	1.689	1.683	29.8	1.525	1.485	29.1	2.296	2.235	26.9
\hat{t}_{Reg}	0.734	0.697	72.2	0.744	0.702	61.5	1.445	1.391	42.6
\hat{t}_{Reg1}	0.319	0.293	34.0	0.431	0.402	32.7	1.201	1.130	29.3
\hat{t}_{Naive}	0.918	0.911	43.3	0.784	0.766	43.5	0.983	0.958	44.2
\hat{t}_{Expnp}^*	0.869	1.403	32.0	0.804	1.173	32.3	1.097	1.322	35.4
$\hat{t}_{\text{Expnpr}}^*$	0.814	1.291	35.1	0.746	1.089	35.2	1.046	1.285	37.1
$\hat{t}_{\text{Regnpr}}^*$	0.700	0.627	73.9	0.658	0.588	66.6	1.033	0.955	50.5
$\hat{t}_{\text{Reg1npr}}^*$	0.294	0.259	36.7	0.355	0.315	37.6	0.836	0.751	37.1

Table 5.3

The Values of $\text{RE}(\hat{t})$ and $\text{RE}(\hat{V}(\hat{t}))$

	y_1		y_2		y_3	
\hat{t}_{Exp}	-0.0024	-0.0015	-0.0040	-0.0242	-0.0045	-0.0267
\hat{t}_{Reg}	-0.0014	-0.0510	-0.0015	-0.0556	-0.0024	-0.0373
\hat{t}_{Reg1}	-0.0008	-0.0812	-0.0009	-0.0684	-0.0017	-0.0596
\hat{t}_{Naive}	-0.1377	-0.0083	-0.1474	-0.0230	-0.1787	-0.0260
\hat{t}_{Expnp}^*	-0.0466	0.6141	-0.0591	0.4582	-0.0988	0.2046
$\hat{t}_{\text{Expnpr}}^*$	-0.0406	0.5860	-0.0540	0.4591	-0.0917	0.2282
$\hat{t}_{\text{Regnpr}}^*$	-0.0198	-0.1038	-0.0315	-0.1077	-0.0720	-0.0752
$\hat{t}_{\text{Reg1npr}}^*$	-0.0193	-0.1191	-0.0310	-0.1124	-0.0713	-0.1015

Table 5.4

The Levels $P_o(\hat{t})$ at 90%, 95% and the $\text{RB}(\hat{t})$

	y_1			y_2			y_3		
\hat{t}_{Exp}	0.873	0.922	0.027	0.870	0.914	0.042	0.852	0.904	0.037
\hat{t}_{Reg}	0.881	0.929	0.024	0.876	0.929	0.022	0.870	0.917	0.025
\hat{t}_{Reg1}	0.866	0.926	0.021	0.873	0.923	0.018	0.860	0.914	0.019
\hat{t}_{Naive}	0.322	0.427	2.126	0.298	0.405	2.187	0.287	0.389	2.239
\hat{t}_{Expnp}^*	0.851	0.906	0.740	0.800	0.874	0.866	0.667	0.758	1.172
$\hat{t}_{\text{Expnpr}}^*$	0.872	0.925	0.666	0.830	0.893	0.820	0.700	0.789	1.114
$\hat{t}_{\text{Regnpr}}^*$	0.839	0.908	0.350	0.806	0.878	0.510	0.712	0.789	0.880
$\hat{t}_{\text{Reg1npr}}^*$	0.804	0.871	0.526	0.767	0.844	0.683	0.678	0.763	0.969

II. The Naive Estimator

The naive estimator registers absolute values of $\text{IB}(\hat{t})$ and $\text{RE}(\hat{t})$ that are very high in relation to the other estimators (Tables 5.1 and 5.3). The same is true for the values of $\text{MSE}(\hat{t})$ (Table 5.1). The values of the observed recovery rates $P_o(\hat{t})$ as well as those of the relative bias $\text{RB}(\hat{t})$ are hardly surprising, considering the size of the point estimate bias (Table 5.4).

The behaviour, in terms of variance and variance estimator (Table 5.2) of \hat{t}_{Naive} , is due to the fact that it constitutes a particular case of \hat{t}_{Exp} , assuming uniform response mechanisms. In a sense, this amounts to assuming that the data are missing randomly.

III. The Adjusted Estimators

The reduction of the bias and the mean square error resulting from the use of the adjusted estimators (Table 5.1) is quite significant, in comparison with the naive estimator, especially for the regression estimators (the estimators \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$). In terms of variance (Table 5.2), we have the following inequalities:

$$V(\hat{t}_{\text{Reglnp}}^*) < V(\hat{t}_{\text{Regnp}}^*) < V(\hat{t}_{\text{Expnp}}^*) < V(\hat{t}_{\text{ExpnpX}}^*),$$

which are analytically difficult to demonstrate. Little variation [in terms of $V(\hat{t})$ and $\mathbb{E}(\hat{V}(\hat{t}))$] is observed between items y_1 and y_2 in light of the little variation between the correlations (0.05). On the other hand, the effect of the correlation with the auxiliary variable on $V(\hat{t})$ and of $\mathbb{E}(\hat{V}(\hat{t}))$ may be observed by comparing items y_1 and y_3 , then y_2 and y_3 : the variations between the correlations are greater in these two cases (0.20 and 0.15 respectively).

In terms of variance estimators (Table 5.2), we observe that:

$$\hat{V}(\hat{t}_{\text{Reglnp}}^*) < \hat{V}(\hat{t}_{\text{Regnp}}^*) < \hat{V}(\hat{t}_{\text{Expnp}}^*),$$

as such is the case for the estimators \hat{t}_{Reg} , \hat{t}_{Regl} and \hat{t}_{Exp} . What is surprising, and is of course due to the effect of the auxiliary variables on the variance components relative to the response mechanisms, is the fact that the estimators \hat{t}_{Expnp}^* overestimate the variance with very large absolute values of $\text{RE}(\hat{V}(\hat{t}))$, while the regression estimators \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$ underestimate the variance with absolute values of $\text{RE}(\hat{V}(\hat{t}))$ that are smaller in relation to those of \hat{t}_{Expnp}^* (Table 5.3). For the estimators \hat{t}_{Expnp}^* , not only is the total variance high in relation to that of the regression estimators, but also the relative contribution of the sampling variance is low (Table 5.2).

In terms of recovery rate (Table 5.4), the estimators \hat{t}_{Expnp}^* yield observed rates that are closer to theoretical rates than the estimators \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$. However, the values of the relative bias $\text{RB}(\hat{t})$ are higher for \hat{t}_{Expnp}^* than for \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$, which makes the confidence intervals less reliable.

IN CONCLUSION

(i) If the goal of the estimation is to reduce bias and mean square error, all the estimators adjusted for non-response perform well in relation to the uniform response

mechanism (which basically amounts to doing nothing about nonresponse). The rate of reduction of the bias of each estimator in relation to the naive estimator is at least 66%. The regression estimators \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$ are the most promising of the various estimators considered (Table 5.1).

(ii) If the goal is to construct confidence intervals, we need a pair of estimators $[\hat{t}, \hat{V}(\hat{t})]$ that simultaneously minimize the absolute biases $|\text{B}(\hat{t})|$ and $|\text{B}(\hat{V}(\hat{t}))|$. Tables 5.1 and 5.2 clearly show that the estimators \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$ are the best. These estimators are less sensitive to nonresponse if we consider the values of $\text{RE}(\hat{t})$ and $\text{RE}(\hat{V}(\hat{t}))$ (Table 5.3). Nevertheless the criterion of reliability of the confidence intervals ($\text{RB}(\hat{t}) < 0.10$) is never met (Table 5.4).

(iii) The behaviour of the estimators adjusted (i) for item y_1 , which is the item the most highly correlated with the auxiliary variable, compared to item y_3 , then (ii) for item y_2 compared to item y_3 (y_3 being the item that is least correlated with the auxiliary variable), shows that with very strong explanatory variables (for y_q and for Θ_{qk}), better results can be achieved not only in terms of less bias $|\text{B}(\hat{t})|$ and $|\text{B}(\hat{V}(\hat{t}))|$ but also in terms of less mean square error (a gain in precision in relation to the naive estimator) and a better recovery rate for the confidence intervals (Tables 5.1 to 5.4).

(iv) The behaviour of the estimators \hat{t}_{Regnp}^* and $\hat{t}_{\text{Reglnp}}^*$, in terms of bias, variance and variance estimation, is consistent with the studies conducted by Särndal and Hui (1981), Särndal and Swenson (1985, 1987), Bethlehem (1988) and Kott (1987) on the usefulness of regression estimators in nonresponse situations and the importance of having good predictor variables for the items of interest and the response mechanisms.

ACKNOWLEDGEMENTS

I wish to express my thanks to Carl-Erik Särndal for his support in every sense of the word in the writing of my Ph.D. thesis, on which this article is based. Despite his many responsibilities and the other demands on his time, he taught me a great deal in this field of sampling, which he masters so well and in which he has become a figure of international prominence through his many published works (articles and books) and collaborative efforts.

I would also like to thank the referees and the Associate Editor for their constructive comments. On the one hand, their observations and suggestions improved the original version of this article. On the other, they provided ideas for subsequent studies.

REFERENCES

- BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- CHICOINEAU, F., PAYEN, J.F., and THÉLOT, C. (1985). Modélisation et redressement des non-réponses: le cas du salaire. *Bulletin of the International Statistical Institute*, LI-3, 15.3, 1-23.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3rd Ed.). New York: Wiley.
- GIOMMI, A. (1985). On the estimation of the individual response probabilities. *Bulletin of the International Statistical Institute*, 2, 577-578.
- GIOMMI, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology*, 13, 127-134.
- GROSBAS, J.-J. (1987b). Les réponses manquantes. In *Les sondages*. (Eds. J.-J. Driesbeke, B. Fichet and F. Tassi). Paris: Economica.
- JOHNSON, N.L., and KOTZ, S. (1970). *Continuous univariate distributions-I*. New York: Houghton.
- KOTT, P.S. (1987). Nonresponse in a periodic sample survey. *Journal of Business and Economic Statistics*, 5, 287-293.
- KRAFT, C.H., LEPAGE, Y., and VAN EEDEN, C. (1983). Some finite-sample size properties of Rosenblatt density estimates. *The Canadian Journal of Statistics*, 11, 95-104.
- OH, H.L., and SCHEUREN, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin and D.B. Rubin), 2, 143-184. New York: Academic Press.
- RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.
- ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of the density function. *Annals of Mathematical Statistics*, 27, 832-837.
- SÄRNDAL, C.-E., and HUI, T-K. (1981). Estimation for non-response situations: to what extent must we rely on models? In *Current Topics in Survey Sampling*. (Eds. D. Krewski, R. Platek and J.N.K. Rao), 227-246. New York: Academic Press.
- SÄRNDAL, C.-E., and SWENSON, B. (1985). Incorporating nonresponse modelling in a general randomization theory approach. *Bulletin of the International Statistical Institute*, LI-3, 15.2, 1-16.
- SÄRNDAL, C.-E., and SWENSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- WEGMAN, E.J. (1972a). Nonparametric probability density estimation: A summary of available methods. *Technometrics*, 14, 533-546.
- WEGMAN, E.J. (1972b). Nonparametric probability density estimation: A comparison of density estimation methods. *Journal of Statistical Computations and Simulations*, 1, 225-245.

Competitors to Genuine π ps Sample Designs: A Comparison

OLIVER SCHABENBERGER and TIMOTHY G. GREGOIRE¹

ABSTRACT

Without-replacement list sampling with probability proportional to some measure of element size has not enjoyed much application in forestry because of the difficulty of implementing such sample strategies, that have been termed π ps designs to distinguish without-replacement sampling from the well-known with-replacement pps designs. In this contribution, an exact π ps strategy (Sunter's variant 2), an approximate π ps design (Sunter's variant 1) and the Rao-Hartley-Cochran random group method are examined and the variances of the respective estimators for total bole volume are computed for four tree populations. The results indicate that compared to the Rao-Hartley-Cochran design Sunter's variant 1 in general leads to higher precision if the relationship between auxiliary information x_k and target characteristic y_k is loose but is sensitive to the ordering of the sampling frame, whereas the Rao-Hartley-Cochran design does not require the sampling frame to be ordered at all and appears to be superior if strong linear relationships between x_k and y_k are present.

KEY WORDS: Probability proportional to size sampling; Fixed sample size; Approximate π ps designs; Empirical comparison.

1. INTRODUCTION

Rao (1978) classifies methods for unequal probability sampling without replacement in two broad categories, (i) sampling schemes, where the inclusion probabilities π_k are proportional to the characteristic of interest, y_k , and the Horvitz-Thompson π estimator \hat{t}_π is utilized; (ii) schemes that entertain statistics other than the Horvitz-Thompson estimator. Strategies in (i) are termed IPPS (inclusion probability proportional to size) and members of (ii) non-IPPS designs. In recent literature, *e.g.*, Särndal *et al.* (1992), selection probabilities when sampling with-replacement are denoted p , whereas their counterparts when sampling without replacement are denoted π . We therefore call sampling designs in (i) genuine π ps strategies in this paper. Both, IPPS and non-IPPS designs have in common, that under exact proportionality, *i.e.*, $\pi_k \propto y_k$ and $n(s) \equiv n \{\text{constant}\}$, it is implied that $\text{Var}(\hat{t}) \equiv 0$ where \hat{t} is the respective estimator used. For this reason, it seems appealing to draw a sample without replacement where $\pi_k \propto y_k$ and to keep the sample size fixed at the same time. Our interest in these methods concerns their utility to sampling needs in forestry.

Several exact π ps designs are available, Rao (1978) gives an in depth account and discussion. Their implementation however is often a non-trivial task and numerically cumbersome for sample sizes usually encountered in forestry practice. Many of these exact π ps strategies require enumeration of all possible samples or use algorithms that become increasingly prohibitive as n increases.

A simple design, which is feasible for $n \leq 10$ is described by Sampford (1967).

In forestry, however, the number of samples to be drawn at any stage of a survey is oftentimes much larger, even after stratification. Consequently, one either approximates the π ps selection process in a manner that allows the inclusion probabilities to be computed exactly, or approximates second-order inclusion probabilities π_{kl} in a design that ensures an exact π ps selection. Rao, Hartley and Cochran (1962) described a non-IPPS design, also known as the random group method, that has gained considerable attention (see also Rao 1966, 1978). It is not a π ps design, since it utilizes an estimator other than \hat{t}_π to ensure zero variance when the π_k are proportional to y_k , but is of remarkable simplicity. An approximate π ps design of the first kind is Sunter's method (Sunter 1977a, 1977b). These two designs are referred to in what follows as RHC and SUN1. Sunter (1986, 1989) described an exact π ps strategy that can be applied if certain stipulated conditions about the ordering of the sampling frame are met and the possible samples can be enumerated to obtain π_{kl} for some pairs of elements. To avoid enumeration we use an approximation to these π_{kl} . This scheme will be called variant 2 or SUN2 in what follows.

Särndal *et al.* (1992) describe the SUN1 and RHC strategies as entailing some loss of efficiency compared to corresponding π ps designs, but no assessment of their comparative efficiency is provided. To our knowledge, none is extant; yet in light of the practical advantages offered by these designs, a comparative assessment would be helpful.

¹ Oliver Schabenberger and Timothy G. Gregoire, Department of Forestry, Section Forest Biometrics, College of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, U.S.A.

The purpose of this study is to compare the performance of the three strategies empirically, using data from forestry field studies and sampling intensities up to 10% which involve reasonably large samples.

The designs SUN1, SUN2, and RHC are appropriate if one has access to a list of population elements from which the sample can be drawn. A complete enumeration of the target characteristic y_k is not anticipated, but the probabilities of inclusion may be made proportional to an auxiliary variable x_k . That is, having complete knowledge about x_k prior to sampling, where it is surmised that x_k is roughly proportional to y_k , we try to achieve $\pi_k \propto x_k$ while $n \equiv \text{constant}$.

In forestry such auxiliary information oftentimes is an easily obtainable characteristic of tree size such as height h , diameter at breast height d , or a combination thereof, which can be used to sample efficiently for bole volume or biomass, y . For example, the geometry of tree stems suggests relationships between d , h , and the volume contained in the tree bole that can be exploited in sampling. In the present investigation, the target parameter is the total bole volume per unit area or in an entire forest stand. In practice, some form of multistage sampling would be used, but for sake of exposition the present comparison includes single stage sampling only.

For the RHC and SUN designs, the auxiliary variables d , d^2 , d^2h and the tree sequence number were used. The sequence number was chosen as an auxiliary variable since in the absence of ordering by size it is clearly unrelated to the target characteristic. It should indicate the sensitivity of competing strategies to uninformative auxiliary information (cf. Rao 1966).

All designs were investigated with samples of intensity 1%, 2%, 5%, and 10%. The performance of the different sampling designs was gauged in terms of the variance of each estimator of $t = \sum_{k=1}^N y_k$. Ratio-of-means estimation following simple random sampling was used as a benchmark, since it utilizes the same auxiliary information. The variances of the sample designs described in the following section were compared to the mean square error of the ratio-of-means estimator (ROM), evaluated using the second order delta method approximation in Sukhatme *et al.* (1984).

2. SAMPLE DESIGNS

2.1 Sunter's Design, Variant 1

Sunter initially proposed two different approximate π ps designs: one relaxes the requirement of proportionality of inclusion probabilities π_k for a subset of the population, the other allows for some variation in sample size (Sunter 1977a, 1977b; Schreuder *et al.* 1990). In order that precision not be unduly sacrificed, it is assumed in the latter case that the variance of $n(s)$ is small, while in the first

case that altering some π_k is not too serious. In this study only the first method was used since the RHC design operates with fixed sample size, too, and it is the comparative feasibility of the Sunter and RHC designs that prompted this study. Särndal *et al.* (1992) describe the allocation of the sample and the computation of the inclusion probabilities in detail. For part of the population, $\pi_k \propto x_k$ where x_k is the auxiliary information available for the k -th subject (or record). Let k^* denote an element in the ordered population. Then for all elements where $k < k^*$ selection is carried out proportional to x_k . The process ends if a total sample of size n is allocated or if $k = k^* = \min\{\min\{k: nx_k/t_k \geq 1\}, N - n + 1\}$ where $t_k = \sum_{j \geq k} x_j$. In the latter case, the remaining samples are selected according to the list-sequential scheme of Bebbington (1975) among those elements for which $k \geq k^*$. As Sunter points out, this sampling scheme has the advantage that only one pass through the sampling frame is necessary. Moreover, the first and second order inclusion probabilities can be computed during this pass through the file. Since the design ensures that $\pi_{kl} > 0 \forall k, l$; $\pi_k \pi_l - \pi_{kl} > 0 \forall k, l$ and n is fixed, the non-negative Yates-Grundy estimator of variance can be readily computed. The first order inclusion probabilities are obtained as $\pi_k = nx_k/T_N$ if $k < k^*$ and $\pi_k = n\bar{x}_k/T_N$ if $k \geq k^*$ where $T_N = \sum_{k=1}^N x_k$ and $\bar{x}_{k^*} = t_{k^*}/(N - k^* + 1)$. Expressions for the second order inclusion probabilities are given in Särndal *et al.* (1992).

Consequently, the ordering of the population affects the performance of the SUN1 design, since the inclusion probabilities and therefore the variance depend on k^* (see (2) below). For large sample sizes the condition $k^* = \min\{\min\{k: nx_k/t_k \geq 1\}, N - n + 1\}$ may be resolved in favor of $k^* = \min\{k: nx_k/t_k \geq 1\}$, which in turn may lead to a premature switch from π ps to SRS sampling owing to the ordering of the sampling frame. Note that $x_k/t_k < x_{k'}/t_{k'}$ for $k' > k$ need not be true since if $x_k > x_{k+1}$ and $t_k > t_{k+1}$ it may well be that x_k/t_k is greater or smaller than x_{k+1}/t_{k+1} . It thus can happen that $nx_k > t_k$ and $nx_{k'} < t_{k'}$, for some k, k' where $k' > k$. In this case, that may occur rather frequently, it is unclear if the switch from π ps to SRS should take place the first time $nx_k \geq t_k$ or not. Sometimes it may happen that for the first two or three elements of the population $nx_k \geq t_k$ but falls below t_k for the main portion of the sampling frame. This is especially the case when n is large and a few very big x_k appear on top of the population list. To stick to Sunter's rule in such a case would in essence be equivalent to drawing a simple random sample.

The π estimator for the population total can be computed as

$$\hat{t}_{\pi \text{SUN1}} = \sum_{k=1}^N \frac{y_k}{\pi_k} I_k, \quad (1)$$

where I_k is the sample inclusion indicator function. The variance is obtained as

$$\text{Var}(\hat{t}_{\pi\text{SUN1}}) = -\frac{1}{2} \sum_{k=1}^N \sum_{l=1}^N \text{Cov}(I_k, I_l) \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2, \quad (2)$$

which is the Yates-Grundy form with $\text{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$ (Särndal *et al.* 1992). We use the notation VAR_{SUN1} for (2) subsequently.

2.2 Sunter's Variant 2

In Sunter (1986, 1989) an exact π ps design is described for samples of size $n > 2$. To fix ideas let $z_k = x_k/T_N$ and order the population such that

$$nz_k < Z_k, \quad k = 1, \dots, N - (n + 1)$$

$$(n - k)z_1 < Z_k, \quad l \geq k \geq N - n,$$

where $Z_k = \sum_{i=k}^N z_i$. Let m_k denote the number of samples out of n still to be drawn when arriving at the k -th population element u_k . Given that the two conditions are met, the following algorithm selects an exact π ps sample. For u_k , $P(u_k | m_k) = nz_k/Z_k$ until $m_k = 0$ or $m_k = N - k$; in the latter case discard one of the remaining units with probability $1 - (m_k z_l/Z_k)$ and retain the others.

It is not always possible to order the population such that the above conditions are met. Sunter (1986) describes an algorithm that checks, whether the ordering is possible. The inclusion probabilities are

$$\begin{aligned} \pi_k &= nz_k \\ \pi_{kl} &= n(n - 1)z_k z_l \gamma_k \quad k \leq N - n - 1, \quad l > k, \end{aligned} \quad (3)$$

where

$$\begin{aligned} \gamma_k &= \frac{1}{Z_{k+1}} \left(1 - \frac{z_1}{Z_2} \right) \dots \left(1 - \frac{z_{k-1}}{Z_k} \right), \\ k &= 2, \dots, N - (n + 1). \end{aligned}$$

The remaining second-order inclusion probabilities, namely π_{kl} for $l > k > N - n$ have to be obtained from enumeration of possible samples which is likely to be infeasible. Sunter argues that (3) gives a good approximation for those pairs of elements, and this approximation has been used here. With these inclusion probabilities, $\hat{t}_{\pi\text{SUN2}}$ is indicated by the right-hand-side (rhs) of (1). An approximation to $\text{Var}(\hat{t}_{\pi\text{SUN2}})$ is given by (2), wherein (3) is used to obtain π_{kl} for $l > k > N - n$.

The differences between SUN1 and SUN2 are noteworthy. With SUN1 the joint inclusion probabilities are computed exactly for all pairs, but the selection is not

genuine π ps because of the introduction of SRS in part. In Sunter's variant 2 the selection is exactly π ps, but $\text{Var}(\hat{t}_{\pi\text{SUN2}})$ can only be approximated. We use VAR_{SUN2} to denote this approximation.

2.3 RHC Design

A description of the RHC design is straightforward; properties of the RHC estimator are well documented in Rao, Hartley and Cochran (1962), and Rao (1966, 1978). After fixing the sample size n , the universe of size N is randomly divided into n groups of size N_i where $N = \sum_i N_i$ ($i = 1, \dots, n$). Let X_{ik} denote auxiliary information for element u_k in group i , $k = 1, \dots, N_i$, and put $X_i = \sum_{k=1}^{N_i} X_{ik}$. From each group one element is selected with selection probability $p_{ik} = X_{ik}/X_i$. The estimator for the total in group i is given as

$$\hat{t}_{i\pi} = \sum_{k=1}^{N_i} \frac{y_{ik}}{p_{ik}} I_{ik},$$

where I_{ik} is the sample inclusion indicator function for element u_k in group i . The population total is then estimated by

$$\hat{t}_{gr} = \sum_{i=1}^n \hat{t}_{i\pi}, \quad (4)$$

with variance

$$\begin{aligned} \text{Var}(\hat{t}_{gr}) &= \frac{1}{N(N - 1)} \left(\sum_{i=1}^n N_i^2 - N \right) \\ &\quad \left(\sum_{k=1}^N T_N y_k^2 / x_k - t^2 \right). \end{aligned} \quad (5)$$

Note that (5) depends on the group sizes and is minimized when all are equal. In our application, we determined N_i such that some groups were of size $N_i = [N/n]_{\text{gif}}$ where gif denotes the greatest integer function and the remainder of size $N_i = [N/n]_{\text{gif}} + 1$. The number of groups of each size is chosen so that the sum of the group sizes is N . If N/n is an integer, all groups are of course of equal size. We denote (5) by VAR_{RHC} in the sequel.

The RHC design is not an exact π ps design, since the subdivision of the population introduces a source of randomness unrelated to the size of the auxiliary variable and (4) is not a Horvitz-Thompson estimator. The inclusion probability depends jointly on the size of X_{ik} and on the probability of an element being assigned to group i . Ordering of the population has no effect on VAR_{RHC} .

3. TREE POPULATIONS

Table 1 shows the tree populations under consideration and Figure 1 displays the relationship between the various choices for x_k and the target characteristic for the yellow poplar population. We notice almost perfect proportionality between d^2h and volume, the relationship between d and volume is clearly curvilinear, and the relationship

between d^2 and volume is intermediate. No noticeable trend between sequence number and volume is apparent in the unordered sampling frame. For the remaining three populations similar patterns hold.

For the four populations and the various combinations of auxiliary variable and sampling intensity, there were no observations for which $nx_k > T_N$, thus no records were measured with certainty.

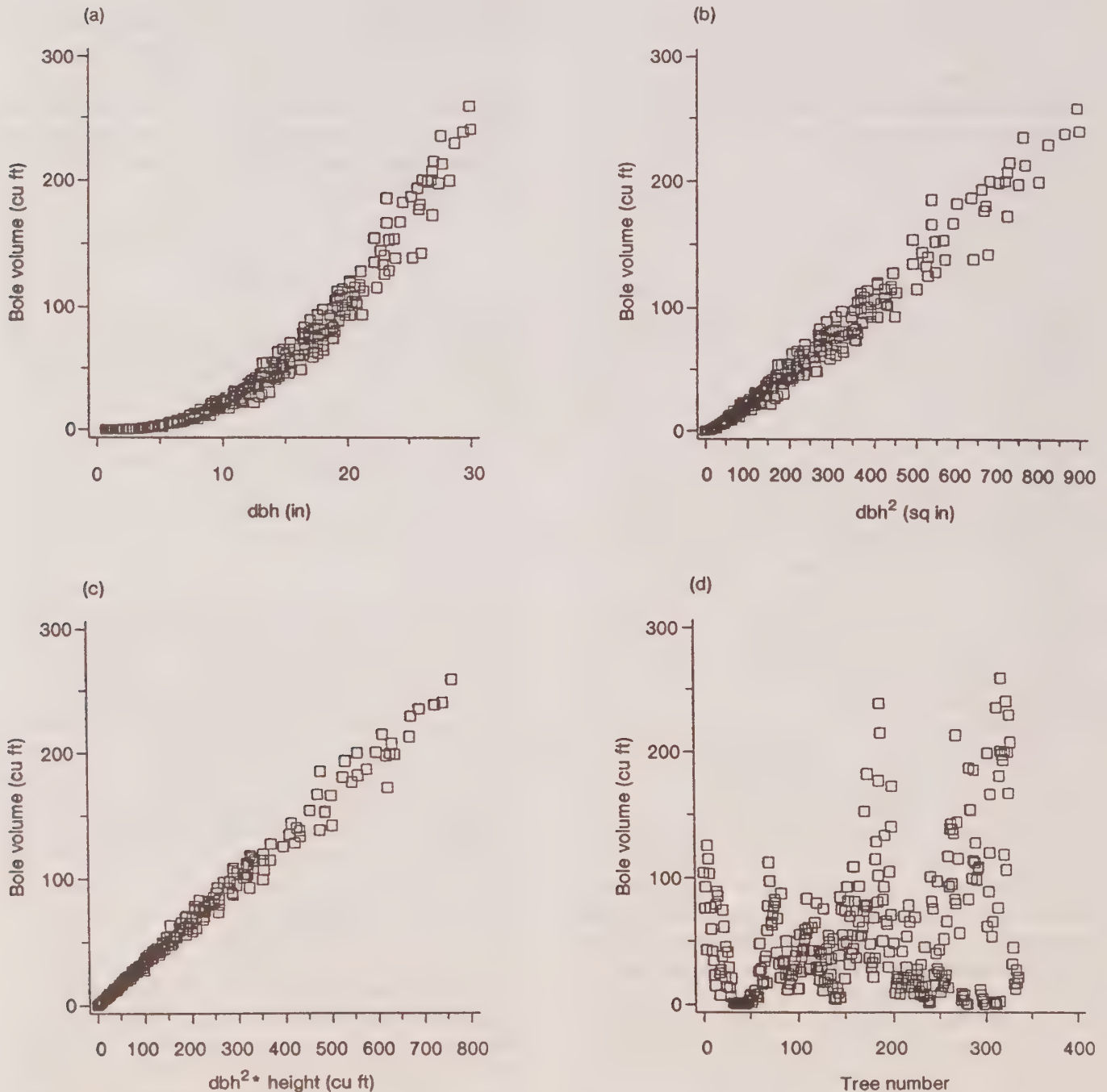


Figure 1. Relation of bole volume to bole dimensions in yellow poplar: (a) diameter at breast height; (b) diameter squared; (c) squared diameter times height; (d) tree sequence number.

Table 1

Tree Populations Examined in an Empirical Comparison of SUN1, SUN2, and RHC

The Last Four Columns Contain Pearson Correlation Coefficients Between x_k and y_k

Species		$N^{(1)}$	$t(ft^3)^{(2)}$	$\rho(y;x)$			
				d	d^2	d^2h	No
Ponderosa pine	<i>Pinus ponderosa</i>	140	9,366.6	0.99	0.99	0.99	0.31
Yellow poplar	<i>Liriodendron tulipifera</i>	336	18,255.5	0.96	0.96	0.99	-0.07
Loblolly pine	<i>Pinus taeda</i>	437	1,835.8	0.96	0.96	0.99	-0.32
Red pine	<i>Pinus resinosa</i>	91	4,075.7	0.96	0.96	0.97	-0.05

(1) N is the number of trees in the population.
(2) t is total volume.

4. RESULTS

4.1 Comparison of Variances

The variance of the estimators of t corresponding to the SUN1, SUN2, and RHC design, expressed as a proportion of the MSE under the ROM strategy are compared in Table 2 for the yellow poplar population for each of the sampling intensities investigated and Table 3 depicts pertinent results for the remaining populations. For the SUN1 strategy, the populations were ordered by decreasing size of X , as recommended by Sunter (1977a, 1977b). We focus initially on the results for the yellow poplar population in Table 2.

Table 2

Relative Performances of SUN1, SUN2 and RHC Design for the Yellow Poplar Population where Ratio-of-means Estimation (ROM) Serves as a Benchmark

$n/N\%$	X	n	$\frac{VAR_{SUN2}}{MSE_{ROM}}$	$\frac{VAR_{SUN1}}{MSE_{ROM}}$	$\frac{VAR_{RHC}}{MSE_{ROM}}$	k^{*1}
1	No	4	4.8120	3.3136	4.7767	
1	d	4	0.6735	0.6684	0.6731	332
1	d^2	4	0.4605	0.4596	0.4613	333
1	d^2h	4	0.3361	0.3378	0.3402	330
2	No	7	5.1327	2.6346	5.0568	
2	d	7	0.7090	0.6982	0.7081	325
2	d^2	7	0.5731	0.5694	0.5751	318
2	d^2h	7	0.4263	0.4542	0.4369	316
5	No	17	5.4938	1.6643	5.2793	
5	d	17	0.7305	0.7808	0.7283	309
5	d^2	17	0.6541	0.6992	0.6608	291
5	d^2h	17	0.4603	1.2638	0.4935	285
10	No	34	5.8326	1.0985	5.3594	
10	d	34	0.7385	0.7083	0.7339	247
10	d^2	34	0.6712	0.9687	0.6864	260
10	d^2h	34	0.4298	3.0140	0.5037	250

¹ k^* is the observation in the ordered sampling frame at which the SUN1 design switches from π ps to SRS sampling.

For a given sampling intensity the precision of all designs relative to ROM increases in the order $X \equiv No, d, d^2, d^2h$; i.e., with increasing proportionality between auxiliary variable and tree bole volume. Given that the approximation of the variance of SUN2 performs well, VAR_{SUN2} can be regarded as measuring the closeness of the RHC and SUN1 designs to matching the efficiency of a genuine π ps selection. At low sampling intensities and with meaningful auxiliary information the two designs do not deviate much from SUN2. The performance of both RHC and SUN1 appears to deteriorate at higher sampling intensities relative to SUN2 depending on the choice of size measure. For $X \equiv d^2h$, in which case $\rho(y;x) \cong 0.99$ (see Table 1), RHC is still .85 (.4298/.5037) as efficient as SUN2 but SUN1 is only .14 (.4298/3.014) as efficient, when $n/N\% = 10$. The performance of RHC and SUN1 relative to SUN2 improves for other choices of X which are less well correlated with Y . Indeed, when $X = No$, SUN1 is much more efficient than SUN2.

A puzzling aspect of these results is the indication that SUN2 is less efficient than either RHC or SUN1 for some choices of auxiliary variable and sampling intensity. We speculate that it may be an artifact of the approximation of some second-order inclusion probabilities incorporated into VAR_{SUN2} . It also may depend on the particular ordering used in SUN1 or the group sizes used in RHC sampling, respectively. It is feasible to calculate the exact $Var(\hat{t}_{\pi SUN2})$ for $n = 2$. We did so for the ponderosa pine and the red pine populations. The results indicate that VAR_{SUN2} approximates the precision of the SUN2 design very well, but is slightly conservative. The ratios $Var(\hat{t}_{\pi SUN2})/VAR_{SUN2}$ took on values between 0.975 and 0.999. For larger sample sizes there is no feasible way to determine how well the approximation VAR_{SUN2} performs.

We focus now on the comparison of RHC to SUN1, again with reference to Table 2. At low sampling intensities, VAR_{SUN1} and VAR_{RHC} are essentially equivalent when $X \equiv d^2h$. But using this auxiliary variable at higher intensities led to a substantially better performance of \hat{t}_{gr} in some cases. The most noteworthy case is $n/N\% = 10$ where \hat{t}_{gr} is nearly 6 times more precise than $\hat{t}_{\pi SUN1}$.

We surmise from these results that the better $x_k \propto y_k$ holds, the better is the precision of \hat{t}_{gr} relative to $\hat{t}_{\pi\text{SUN}}$ owing chiefly to the effect of k^* on VAR_{SUN} . Small values of k^* indicate an early switch to a SRS selection and coincide with small values of $\text{VAR}_{\text{SUN2}}/\text{VAR}_{\text{SUN1}}$. Large values of k^* on the other hand correspond to variance ratios close to 1. For yellow poplar, $n/N\% = 10$ and $X \equiv d^2h$ the SUN1 design selects only three-fourths of the population according to a π ps design; we conjecture that the early transition to SRS serves also as an explanation for its poor performance compared to the RHC design. When $X = \text{tree sequence number}$, SUN1 is much more precise than RHC, and its relative precision increases as n increases.

The sharp improvement in efficiency when using an auxiliary variable other than tree sequence number provides an indication of the effectiveness of the strategies discussed here when X is positively correlated to Y , and to the liability of sampling with probability proportional to an auxiliary variable when it is unrelated to Y .

The pattern evident in the results for yellow poplar are generally seen, also, in the results for the other species. Some of them are summarized in Table 3. For ponderosa pine SUN1 relative to RHC is always less precise when $X \equiv d^2h$ regardless of the sampling intensity and SUN2 performs always best when this variable is used. For all species the combination $n/N\% = 10$, $X \equiv d^2h$ leads to low precision of SUN1 compared to the other designs and with the exception of the loblolly pine population, SUN1 performs poorer than ratio-of-means estimation. For all populations, the order of magnitude better precision of ROM over the genuine π ps, non-IPPS or approximate π ps design when $X = \text{tree sequence number}$ is remarkable.

From Figure 1 it can be seen that the ordering of volume by tree numbers is haphazard, *i.e.*, the sequence number carries no information about bole volume. And, there is a price to pay if one uses this uninformative auxiliary information to determine inclusion probabilities. The inefficiency of unequal probability sampling in presence of uninformative auxiliary information is an important limitation for the simultaneous estimation of multiple population attributes, where some may be closely related to the auxiliary design variable but others might be uncorrelated with it. Rao (1966) discusses this point in detail and he proposes alternative estimators based on the unbiased estimators in equal probability sampling and the estimator $\hat{t}_{gr(alt)} = N \sum_i y_i \xi_i$, where $\xi_i = \sum_k^N p_{ik}$ in the RHC design. Applying this estimator in the case of unequal probability sampling leads to bias, but to better mean-square error performance. For the RHC design with $X = \text{tree sequence number}$, the alternative estimator proposed by Rao (1966) improved the ratio $\text{MSE}_{\text{RHC}(alt)}/\text{MSE}_{\text{ROM}}$ remarkably. For the yellow poplar population for example, these ratios were between 1.34 ($n = 4$) and 2.58 ($n = 34$), corresponding

to a mean square error of the alternative estimator of only 28% to 48% ($n = 34$) of the RHC estimator (5). Similar patterns hold for the other tree species.

Since the alternative estimator is inconsistent, its bias does not depend on n , the larger ratios within the range for each species appear for larger sample sizes. It thus seems reasonable to limit the use of this estimator to smaller sample sizes. When n gets larger, another alternative is to use a ratio estimator, *e.g.*, Hajek's estimator $N\{(\sum y_i/\pi_i)/(\sum 1/\pi_i)\}$ under a genuine π ps design.

Table 3
Pertinent Results About the Relative Performances of
SUN1, SUN2 and RHC Design for the Remaining
Populations where Ratio-of-means Estimation (ROM)
Serves as a Benchmark

$n/N\%$	X	n	$\frac{\text{VAR}_{\text{SUN2}}}{\text{MSE}_{\text{ROM}}}$	$\frac{\text{VAR}_{\text{SUN1}}}{\text{MSE}_{\text{ROM}}}$	$\frac{\text{VAR}_{\text{RHC}}}{\text{MSE}_{\text{ROM}}}$	k^*
Ponderosa Pine						
1	No	2	1.9608	1.9794	1.9507	
1	d^2h	2	0.1050	0.1096	0.1077	137
2	No	3	2.2976	1.9264	2.2275	
2	d^2h	3	0.1768	0.1919	0.1859	135
5	No	7	2.8717	2.0681	2.7819	
5	d^2h	7	0.3113	0.3890	0.3670	129
10	No	14	3.2528	2.2745	3.0294	
10	d^2h	14	0.2928	1.3724	0.4488	97
Red Pine ¹						
2	No	2	2.0210	1.9485	2.0029	
2	d^2h	2	0.9076	0.9026	0.9104	90
5	No	5	2.9295	2.3141	2.8236	
5	d^2h	5	0.8874	1.3456	0.8991	87
10	No	9	3.5548	2.0124	3.2958	
10	d^2h	9	0.8699	1.3192	0.8942	81
Loblolly Pine						
1	No	5	4.8011	3.7104	4.7625	
1	d^2h	5	0.4043	0.4161	0.4174	431
2	No	9	5.5940	3.7441	5.5044	
2	d^2h	9	0.5129	0.5510	0.5476	419
5	No	22	6.5290	3.3082	6.5253	
5	d^2h	22	0.5035	0.6385	0.6085	406
10	No	44	7.7977	2.6635	6.5708	
10	d^2h	44	0.3854	0.7214	0.6146	375

¹ The sampling intensity 1% was omitted since it would have resulted in $n = 1$.

4.2 The Effect of Ordering on The Precision of Sunter's Variant 1

Sunter and others have noted that the precision of the SUN1 design depends on the ordering of the population. The recommendation to sort the sampling frame by decreasing size of x_k 's is rooted in the assumption that larger x_k are more likely to be proportional to y_k than smaller ones. The goal is to apply the π ps part of the SUN1 design not only to as big a portion of the population as possible but also to those elements for which $x_k \propto y_k$ holds best. Under this assumption it was thus advised to put the elements with large x_k values at the top of the frame. However, it is clear that this is only a rough rule of thumb, since the assumption of greater proportionality with increasing size may not hold.

To investigate the effect of ordering the ponderosa pine and red pine populations were first sorted by increasing x_k and then grouped into 10 groups of approximately equal size. The Pearson correlation coefficient between x_k and y_k was computed within each group and the populations were then sorted by

- groups of decreasing correlation and increasing size of x_k within each group,
- groups of decreasing correlation and decreasing size of x_k in each group

and SUN1 sampling was repeated for the combinations of x_k 's and sampling intensity 10%. Table 4 shows the results.

Table 4

Var_{SUN1}/MSE_{ROM} for Ponderosa Pine and Red Pine and Different Ways of Ordering the Population

X	Ponderosa Pine Ordered by			Red Pine Ordered by		
	decr. x_k	decr. ρ incr. x_k	decr. ρ decr. x_k	decr. x_k	decr. ρ incr. x_k	decr. ρ decr. x_k
d	0.5614	0.6165	0.6043	1.0307	1.0236	0.6454
d^2	0.3478	0.6562	0.5869	1.2077	0.9373	0.6948
d^2h	1.3724	60.861	0.4459	1.3192	0.8674	0.7461

The results are rather surprising. For red pine the order by decreasing correlation improved all measures of precision. Sorting by increasing x_k within each group now made VAR_{SUN1} very close to VAR_{RHC}, and with $x = d^2h$, VAR_{SUN1} < VAR_{RHC}. Sorting by decreasing x_k within each group achieved an even greater improvement. In contrast to these results, sorting the ponderosa pine population by decreasing ρ and increasing x_k made things worse. The very high value of 60.861 is caused by a premature

switch to SRS, since in this setting k^* is only 28, corresponding to only 20% of the population being sampled π ps. Moreover, using order of decreasing ρ and decreasing x_k improved VAR_{SUN1} only for $x = d^2h$.

These results indicate that there may exist an order that minimizes VAR_{SUN1} and may yield higher precision than a simple ordering by decreasing value of X . But this order will usually differ depending upon the auxiliary information, and even an ordering that is reasonable on intuitive grounds may give unanticipated results. It is not known if any ordering is optimal in the sense of minimizing Var($\hat{t}_{\pi\text{SUN1}}$) for the approximate π ps design used in this study. According to our present knowledge no optimal strategy has been described.

5. DISCUSSION AND CONCLUSION

Employing some meaningful auxiliary information leads to a considerable gain in precision in the unequal probability designs compared to a ratio-of-means estimation.

A choice between the two Sunter designs can be made on grounds of the relationship between size measure and target characteristic. When $X \propto Y$ is strong, SUN2 offers advantage over SUN1, and SUN1 appears preferable when the relationship is weak. Based on our results, the approximate π ps strategy, SUN1 and the non-IPPS design RHC appear to come fairly close to the efficiency offered by genuine π ps selection. With increasing sampling intensity, however, the highest precision is obtained with the SUN2 design. But the quality of the approximation VAR_{SUN2} in this case is unclear.

If one's aim is to use an approximate π ps or a non-IPPS strategy then the RHC design with estimator \hat{t}_{gr} appears to offer advantages over the Sunter design with $\hat{t}_{\pi\text{SUN}}$, at least for the tree populations studied here with the objective of estimating total bole volume. At reasonably low sampling intensities, both estimators appear to be equally precise.

An advantage of the RHC design is its simplicity. An operational advantage is that it can be applied to every population because it is impervious to its ordering and provides an unbiased estimation within each group. While the first criterion is also met by Sunter's variant 1, the ordering there clearly affects the precision of the estimator $\hat{t}_{\pi\text{SUN1}}$. Variant 2 can only be used if some ordering of the population meets the conditions given in Section 2.2. Otherwise the selection algorithm does not produce a sample of exactly size n .

The precision of the RHC method, however, depends on the group sizes employed. The algorithm given in Section 2.3 is optimal.

While a particular ordering may improve the precision of $\hat{t}_{\pi\text{SUN1}}$, it is unclear at present how to discern an optimal ordering and a fixed sample size. Moreover an optimal

ordering of one choice of auxiliary variable or attribute of interest may be deleterious when implemented with a different auxiliary variable or attribute.

All strategies can be disastrous with uninformative auxiliary information.

Finally and to the extent that computational burden is a meaningful criterion, RHC is arguably less burdensome than variant 1 of Sunter's design.

ACKNOWLEDGMENT

We gratefully acknowledge the comments and suggestions by J.N.K. Rao, C.-E. Särndal, and A. Sunter who reviewed earlier versions of the manuscript as well as the helpful comments of the referees whose contribution helped to improve the paper substantially.

REFERENCES

- BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.
- RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā A*, 28, 47-60.
- RAO, J.N.K. (1978). Sampling designs involving unequal probabilities of selection and robust estimation of a finite population total. *Contributions to Survey Sampling and Applied Statistics* (H.A. David, Ed.), New York: Academic Press, 69-86.
- RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society B*, 24, 482-491.
- SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- SÄRNDAL, C.-E., SWENSSON B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SUKHATME, P.V., SUKHATME, S., and ASOK, C. (1984). *Sampling Theory of Surveys with Applications (3rd Ed.)*. Iowa State University Press.
- SUNTER, A. (1977a). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.
- SUNTER, A. (1977b). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.
- SUNTER, A. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Review*, 54, 33-50.
- SUNTER, A. (1989). Updating size measures in a PPSWOR design. *Survey Methodology*, 15, 253-260.
- SCHREUDER, H.T., LI, H.G., and SADOOGHI-ALVANDI, S.M. (1990). Sunter's pps Without Replacement Sampling as an Alternative to Poisson Sampling. USDA Forest Service Research Paper RM-290.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1994. An asterisk indicates that the person served more than once.

- M. Bankier, *Statistics Canada*
 * G.E. Battese, *University of New England, Australia*
 Y. Beaucage, *Statistics Canada*
 * D.R. Bellhouse, *University of Western Ontario*
 N. Bennett, *Yale University*
 J. Bethel, *Westat, Inc.*
 * J. Bethlehem, *Central Bureau of Statistics, The Netherlands*
 P. Biemer, *Research Triangle Institute*
 * D.A. Binder, *Statistics Canada*
 R.L. Chambers, *Australian National University*
 S. Cheung, *Statistics Canada*
 * N. Chinnappa, *Statistics Canada*
 G.H. Choudhry, *Statistics Canada*
 M.P. Cohen, *U.S. National Center for Education Statistics*
 * M.J. Colledge, *Statistics Canada*
 L.H. Cox, *U.S. Environmental Protection Agency*
 C.Z.F. Clark, *U.S. Department of Agriculture*
 R. Cochran, *University of Wyoming*
 F. Conrad, *U.S. Bureau of Labor Statistics*
 N. Cressie, *Iowa State University*
 * J.-C. Deville, *Institut National de la Statistique et des Études Économiques*
 P. Dick, *Statistics Canada*
 D. Dillman, *Washington State University*
 D. Dolson, *Statistics Canada*
 * J.D. Drew, *Statistics Canada*
 * F. Dupont, *Institut National de la Statistique et des Études Économiques*
 W.S. Edwards, *Westat, Inc.*
 E.P. Ericksen, *Temple University*
 R.E. Fay, *U.S. Bureau of the Census*
 * W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistics Canada*
 M. Ghosh, *The University of Florida*
 * M. Gonzalez, *U.S. Office of Management and Budget*
 H. Gough, *Statistics Canada*
 * R.M. Groves, *University of Maryland*
 * J.-P. Gwet, *Statistics Canada*
 K.P. Hapuarachchi, *Statistics Canada*
 H. Hogan, *U.S. Bureau of the Census*
 * D. Holt, *University of Southampton*
 A.Z. Israëls, *Central Bureau of Statistics, The Netherlands*
 R. Jamieson, *Statistics Canada*
 W.D. Kalsbeek, *University of North Carolina – Chapel Hill*
 * G. Kalton, *Westat, Inc.*
 * P.S. Kott, *National Agricultural Statistics Service*
 * J. Kovar, *Statistics Canada*
 * P. Lahiri, *University of Nebraska – Lincoln*
 P. Lavallée, *Statistics Canada*
 H. Lee, *Statistics Canada*
 J.M. Lepkowski, *University of Michigan*
 J.T. Lessler, *Batelle*
 N.Y. Luther, *East-West Center*
 P. Lys, *Statistics Canada*
 D. Malec, *National Centre for Health Statistics*
 * H. Mantel, *Statistics Canada*
 M. March, *Statistics Canada*
 H. Mariotte, *Institut National de la Statistique et des Études Économiques*
 * A. Mason, *East-West Center*
 N. Mathiowetz, *Agency for Health Care and Research*
 P. Miller, *Narthurstan*
 B. Nandram, *Worcester Polytechnic Institute of Mathematical Sciences*
 J. Nealon, *National Agricultural Statistics Service*
 J. Neter, *University of Georgia*
 * D. Pfeffermann, *Hebrew University*
 N.G.N. Prasad, *University of Alberta*
 M. Ramos, *U.S. Bureau of the Census*
 * J.N.K. Rao, *Carleton University*
 * L.-P. Rivest, *Université Laval*
 L. Rizzo, *Westat, Inc.*
 G. Roberts, *Statistics Canada*
 K. Rust, *Westat, Inc.*
 * I. Sande, *Bell Communications Research, U.S.A.*
 * C.-E. Särndal, *Université de Montréal*
 J. Schafer, *Pennsylvania State University*
 * W.L. Schaible, *U.S. Bureau of Labor Statistics*
 * F.J. Scheuren, *George Washington University*
 I. Schiopu-Kratina, *Statistics Canada*
 * J. Sedransk, *State University of New York*
 A.C. Singh, *Statistics Canada*
 * C.J. Skinner, *University of Southampton*
 E.A. Stasny, *The Ohio State University*
 T.W.F. Stroud, *Queen's University*
 C.M. Suchindran, *University of North Carolina*
 J. Tanur, *State University of New York – Stony Brook*
 R. Tourangeau, *National Opinion Research Center*
 R. Treder, *Statistical Sciences Inc.*
 M.E. Thompson, *University of Waterloo*
 J. Tourigny, *Statistics Canada*
 * R. Valliant, *U.S. Bureau of Labor Statistics*
 K.W. Wachter, *University of California – Berkeley*
 * J. Waksberg, *Westat, Inc.*
 M. Weekr, *Research Triangle Institute*
 G.C. White, *Colorado State University*
 W.E. Winkler, *U.S. Bureau of the Census*
 * K.M. Wolter, *National Opinion Research Center*
 * T. Wright, *Oak Ridge National Laboratory*
 E. Zanutto, *Harvard University*
 * A. Zaslavsky, *Harvard University*
 J.V. Zidek, *University of British Columbia*

Acknowledgements are also due to those who assisted during the production of the 1994 issues: S. Beauchamp (Photocomposition) and M. Haight (Translation Services). Finally we wish to acknowledge S. DiLoreto, M.M. Kent, C. Larabie and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 43, No. 4, 1994

	<i>Page</i>
Fully Bayesian approach to image restoration with an application in biogeography <i>J. Heikkinen and H. Högmänder</i>	569
Dose-response models for correlated multinomial data from development toxicity studies <i>Y. Zhu, D. Krewski and W. H. Ross</i>	583
A procedure for estimating the unconditional cumulative incidence curve and its variability for the human immunodeficiency virus <i>J.W. Hay and F.A. Wolak</i>	599
A dynamic changepoint model for detecting the onset of growth in bacteriological infections <i>J. Whittaker and S. Frühwirth-Schnatter</i>	625
Effect of parameter estimation on fertilizer optimization <i>D. Wallach and P. Loisel</i>	641
<i>General Interest Section</i>	
Modelling maximum oxygen uptake – a case-study in non-linear regression model formulation and comparison <i>A.M. Nevill and R.L. Holder</i>	653
<i>Statistical Algorithms</i>	
AS 295 A Federov exchange algorithm for D-optimal design <i>A.J. Miller and K.-K. Nguyen</i>	669
<i>Correction</i>	
Correction to algorithm AS 274: Least squares routines to supplement those of Gentleman <i>A.J. Miller</i>	678
<i>Statistical Software Reviews</i>	
STAT-ITCF	679
<i>Author Index</i>	683
<i>Corrigendum</i>	
Bayesian estimation of the binomial parameter n <i>M.P. Wiper and L.I. Pettit</i>	685

Printed in Great Britain at the Alden Press, Oxford

This journal is printed on acid-free paper

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 43, No. 4, 1994

Page	
569	Fully Bayesian approach to image restoration with an application in biogeography <i>J. Heikkinen and H. Högmänder</i>
583	Dose-response models for correlated multinomial data from development toxicity studies <i>Y. Zhu, D. Krewski and W. H. Ross</i>
599	A procedure for estimating the unconditional cumulative incidence curve and its variability for the human immunodeficiency virus <i>J. W. Hay and F. A. Wolak</i>
625	A dynamic changepoint model for detecting the onset of growth in bacteriological infections <i>J. Whitaker and S. Frühwirth-Schnatter</i>
641	Effect of parameter estimation on fertilizer optimization <i>D. Wallach and P. Loisel</i>
653	General Interest Section Modelling maximum oxygen uptake – a case-study in non-linear regression model formulation and comparison <i>A. M. Nevill and R. L. Holder</i>
669	Statistical Algorithms AS 295 A Federov exchange algorithm for D-optimal design <i>A. J. Miller and K.-K. Nguyen</i>
678	Correction Correction to algorithm AS 274: Least squares routines to supplement those of Gentleman <i>A. J. Miller</i>
679	STAT-ITCF Statistical Software Reviews
683	Author Index
685	Corrigendum Bayesian estimation of the binomial parameter n <i>M. P. Wiper and L. T. Pettit</i>

Printed in Great Britain at the Alden Press, Oxford
This journal is printed on acid-free paper

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1994. Un astérisque indique que la personne a participé plus d'une fois.

- J.T. Lessler, *Batelle*
 N.Y. Luther, *East-West Center*
 P. Lys, *Statistique Canada*
 D. Malec, *National Centre for Health Statistics*
 H. Mantel, *Statistique Canada*
 M. March, *Statistique Canada*
 H. Martotte, *Institut National de la Statistique et des Etudes Economiques*
 A. Mason, *East-West Center*
 N. Mathiowetz, *Agency for Health Care and Research*
 P. Miller, *Narthurstan*
 B. Nandram, *Worcester Polytechnic Institute of Mathematical Sciences*
 J. Nealon, *National Agricultural Statistics Service*
 J. Neter, *University of Georgia*
 D. Pfeffermann, *Hebrew University*
 N.G.N. Prasad, *University of Alberta*
 M. Ramos, *U.S. Bureau of the Census*
 J.N.K. Rao, *Carleton University*
 L.-P. Rivest, *Université Laval*
 L. Rizzo, *Westat, Inc.*
 G. Roberts, *Statistique Canada*
 K. Rust, *Westat, Inc.*
 I. Sande, *Bell Communications Research, U.S.A.*
 C.-E. Sarnad, *Université de Montréal*
 J. Schafer, *Pennsylvania State University*
 W.L. Schaible, *U.S. Bureau of Labor Statistics*
 F.J. Schuren, *George Washington University*
 I. Schiopu-Krathina, *Statistique Canada*
 J. Sedransk, *State University of New York*
 A.C. Singh, *Statistique Canada*
 C.J. Skinner, *University of Southampton*
 E.A. Siasny, *Ohio State University*
 T.W.F. Stroud, *Queen's University*
 C.M. Suchindran, *University of North Carolina*
 J. Tanur, *State University of New York - Stony Brook*
 M.E. Thompson, *University of Waterloo*
 R. Tourangeau, *National Opinion Research Center*
 J. Tourigny, *Statistique Canada*
 R. Tredar, *Statistical Sciences Inc.*
 R. Valliant, *U.S. Bureau of Labor Statistics*
 K.W. Wachter, *University of California - Berkeley*
 J. Wakseberg, *Westat, Inc.*
 M. Weeker, *Research Triangle Institute*
 G.C. White, *Colorado State University*
 W.B. Winkler, *U.S. Bureau of the Census*
 K.M. Wolter, *National Opinion Research Center*
 T. Wright, *Oak Ridge National Laboratory*
 E. Zanutto, *Harvard University*
 A. Zaslavsky, *Harvard University*
 J.V. Zidek, *University of British Columbia*
- M. Bankier, *Statistique Canada*
 G.E. Battese, *University of New England, Australia*
 D.R. Bellhouse, *University of Western Ontario*
 N. Bennett, *Yale University*
 J. Bethel, *Westat, Inc.*
 J. Bethlehem, *Statistics Netherlands*
 P. Biemer, *Research Triangle Institute*
 D.A. Binder, *Statistique Canada*
 R.L. Chambers, *Australian National University*
 S. Cheung, *Statistique Canada*
 N. Chinappa, *Statistique Canada*
 G.H. Choudhry, *Statistique Canada*
 M.P. Cohen, *U.S. National Center for Education Statistics*
 M.J. Colledge, *Statistique Canada*
 L.H. Cox, *U.S. Environmental Protection Agency*
 C.Z.F. Clark, *U.S. Department of Agriculture*
 R. Cochran, *University of Wyoming*
 F. Conrad, *U.S. Bureau of Labor Statistics*
 N. Cressie, *Iowa State University*
 J.-C. Deville, *Institut National de la Statistique et des Etudes Economiques*
 P. Dick, *Statistique Canada*
 D. Dillman, *Washington State University*
 D. Dolson, *Statistique Canada*
 J.D. Drew, *Statistique Canada*
 F. Dupont, *Institut National de la Statistique et des Etudes Economiques*
 W.S. Edwards, *Westat, Inc.*
 E.P. Erickson, *Temple University*
 R.E. Fay, *U.S. Bureau of the Census*
 W.A. Fuller, *Iowa State University*
 J. Gambino, *Statistique Canada*
 M. Ghosh, *The University of Florida*
 M. Gonzalez, *U.S. Office of Management and Budget*
 H. Gough, *Statistique Canada*
 R.M. Groves, *University of Maryland*
 J.-P. Gwet, *Statistique Canada*
 K.P. Hapuarachchi, *Statistique Canada*
 H. Hogan, *U.S. Bureau of the Census*
 D. Holt, *University of Southampton*
 A.Z. Israels, *Netherlands Central Bureau of Statistics*
 R. Jamieson, *Statistique Canada*
 W.D. Kalsbeek, *University of North Carolina - Chapel Hill*
 G. Kalton, *Westat, Inc.*
 P.S. Kott, *National Agricultural Statistics Service*
 J. Kovar, *Statistique Canada*
 P. Lahiri, *University of Nebraska - Lincoln*
 P. Lavallée, *Statistique Canada*
 H. Lee, *Statistique Canada*
 J.M. Lepkowski, *University of Michigan*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1994: S. Beauchamp (Photocomposition) et M. Haight (Services de traduction). Finalement on désire exprimer notre reconnaissance à S. DiLoreto, M.M. Kent, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes-ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

5. ANALYSE ET CONCLUSION

L'utilisation d'une information supplémentaire significative pour effet d'accroître sensiblement la précision des plans d'échantillonnage avec probabilités inégales par rapport à l'estimation du rapport de moyennes.

Le choix de l'un ou l'autre des plans de Sunter dépend de la relation entre la mesure de taille et le caractère étudié. Lorsque la relation $X \propto Y$ est forte, SUN2 est plus avantageux que SUN1; par contre, lorsque la relation est faible, SUN1 semble préférable. D'après les résultats de notre étude, le plan π pt approximatif (SUN1) et le plan non-PSPT (RHC) ont une efficacité qui semble se rapprocher passablement de celle d'un plan π pt authentique. Toutefois, lorsque le taux de sondage augmente, le plan SUN2 est le plus efficace de tous, bien que la qualité de l'approximation VAR_{SUN2} soit incertaine dans ce cas.

Si l'on doit choisir entre un plan π pt approximatif et un plan non-PSPT dans le but d'estimer le volume total de fût, le plan RHC, avec \hat{f}_{gr} comme estimateur, semble plus avantageux que le plan de Sunter avec \hat{f}_{SUN} comme estimateur, du moins pour les populations d'arbres étudiées ici. Lorsque le taux de sondage est relativement faible, les deux estimateurs paraissent aussi précis l'un que l'autre. Un des avantages du plan RHC est sa simplicité. Un autre avantage de ce plan, d'ordre pratique celui-là, est qu'il peut s'appliquer à chaque population parce qu'il est indépendant du classement des éléments de la population et qu'il produit une estimation non biaisée dans chaque groupe. Tandis que la simplicité est aussi une caractéristique de la variante 1 de Sunter, le classement des éléments, dans ce plan, influe nettement sur la précision de l'estimateur \hat{f}_{SUN1} . Quant à la variante 2, elle ne peut être utilisée que si le classement des éléments de la population satisfait les conditions énoncées dans la section 2.2. Autrement, l'algorithme de sélection ne produira pas un échantillon exactement de taille n .

Néanmoins, la précision de la méthode RHC dépend de la taille des groupes. L'algorithme présenté dans la section 2.3 est optimal.

Bien que l'on puisse accroître la précision de \hat{f}_{SUN1} grâce à un mode de classement particulier, on ne sait trop encore comment faire la différence entre le mode de classement optimal et la taille d'échantillon fixe. En outre, il peut être désastreux d'effectuer un classement optimal selon une autre modalité de variable auxiliaire (ou caractéristique) que celle ayant déjà servi à ce classement. Toutes les méthodes peuvent avoir des effets malheureux en présence d'information supplémentaire non informative. Enfin, et dans la mesure où le nombre de calculs à effectuer est un critère important, le plan RHC est, pourrait-on dire, moins lourd que la variante 1 de Sunter.

REMERCIEMENTS

Nous tenons à exprimer notre reconnaissance à MM. J.N.K. Rao, C.-E. Särndal et A. Sunter, qui nous ont communiqué leurs commentaires et leurs suggestions et ont revu des versions antérieures de cet exposé, ainsi qu'aux rapporteurs, qui, par leurs commentaires, ont contribué à en améliorer sensiblement le contenu.

BIBLIOGRAPHIE

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

RAO, J.N.K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhyā A*, 28, 47-60.

RAO, J.N.K. (1978). Sampling designs involving unequal probabilities of selection and robust estimation of a finite population total. *Contributions to Survey Sampling and Applied Statistics* (Eds. H.A. David), New York: Academic Press, 69-86.

RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 24, 482-491.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SUKHATME, P.V., SUKHATME, S., et ASOK, C. (1984). *Sampling Theory of Surveys with Applications* (3ième éd.). Iowa State University Press.

SUNTER, A. (1977a). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A. (1977b). Response burden, sample rotation, and classification renewal in economic surveys. *Revue Internationale de Statistique*, 45, 209-222.

SUNTER, A. (1986). Solutions to the problem of unequal probability sampling without replacement. *Revue Internationale de Statistique*, 54, 33-50.

SUNTER, A. (1989). Mise à jour de la taille de population dans un plan PPTS. *Techniques d'enquête*, 15, 263-270.

SCHREUDER, H.T., LI, H.G., et SADOOGHI-ALVANDI, S.M. (1990). Sunter's pps Without Replacement Sampling as an Alternative to Poisson Sampling. USDA Forest Service Research Paper RM-290.

L'inefficacité de l'échantillonnage avec probabilités inégales attribuable à la présence d'information supplémentaire non informative est un inconvénient majeur lorsqu'on veut estimer simultanément plusieurs caractères de la population, dont certains peuvent être fortement corrélés avec la variable auxiliaire et d'autres, pas du tout. Rao (1966) traite la question en détail et propose des estimateurs de rechange inspirés de l'estimateur non biaisé, dans le cas de l'échantillonnage avec probabilités égales, et de l'estimateur $\hat{y}_{gr}^{(alt)} = N \sum_{i=1}^N y_i \xi_i$, où $\xi_i = \sum_{k=1}^K p_{ik}$, dans le cas du plan RHC. Dans un plan d'échantillonnage avec probabilités inégales, l'estimateur de rechange est biaisé mais son erreur quadratique moyenne est plus faible. Étant donné un plan RHC et $X =$ numéro de séquence de l'arbre, le rapport $EQM_{RHC}^{(alt)} / EQM_{ROM}$ se trouve très amélioré grâce à l'estimateur proposé par Rao (1966). Pour la population de tulipiers d'Amérique par exemple, le rapport en question prend des valeurs allant de 1,34 ($n = 4$) à 2,58 ($n = 34$), ce qui implique que l'erreur quadratique moyenne de l'estimateur de rechange n'est l'équivalent que de 28 à 48% ($n = 34$) de la variance de l'estimateur RHC (5). On peut faire les mêmes observations pour les trois autres essences d'arbre.

Comme l'estimateur de rechange n'est pas convergent, son biais ne dépend pas de n ; pour chaque essence, on observe des ratios plus élevés lorsqu'il s'agit de gros échantillons. Il semble donc raisonnable de limiter l'utilisation de l'estimateur de rechange aux petits échantillons. Lorsque n augmente, on a aussi la possibilité d'utiliser un estimateur par quotient – par exemple, l'estimateur de Hajek $N \{ (\sum y_i / \pi_i) / (\sum 1 / \pi_i) \}$ – suivant un plan π_{pi} authentique.

4.2 Effet du classement des éléments sur la précision de la variante 1 de Sunter

Sunter et d'autres ont remarqué que la précision du plan SUN1 dépendait du classement des éléments de la population. La recommandation selon laquelle on classe les éléments de la base de sondage par ordre décroissant selon la valeur de x_k trouve sa justification dans l'hypothèse que x_k est plus susceptible d'être proportionnel à y_k selon la valeur de x_k est plus susceptible d'être proportionnel à y_k pour des valeurs élevées peut ne pas se vérifier.

Pour analyser l'effet du classement d'éléments, nous avons trié les populations de pins ponderosa et de pins rouges par ordre croissant selon la valeur de x_k , puis nous avons divisé chacune des populations en 10 groupes de taille comparable. Nous avons calculé le coefficient de corrélation de Pearson entre x_k et y_k pour chaque groupe,

puis nous avons classé les éléments de chaque groupe selon deux combinaisons de critères:

(a) corrélation décroissante et valeur de x_k croissante,

(b) corrélation décroissante et valeur de x_k décroissante,

enfin, nous avons répété l'échantillonnage selon SUN1 pour les modalités de x_k et un taux de sondage de 10%. Les résultats figurent au tableau 4.

Tableau 4
 Var_{SUN1} / EQM_{RM1} pour le pin ponderosa et le pin rouge, selon divers modes de classement de la population

Pins ponderosa, classés selon				Pins rouges, classés selon			
X	p décr.	x_k décroiss.	x_k décr.	X	p décr.	x_k décroiss.	x_k décr.
d	0,5614	0,6165	0,6043	1,0307	1,0236	0,6454	
d^2	0,3478	0,6562	0,5869	1,2077	0,9373	0,6948	
d^2h	1,3724	60,861	0,4459	1,3192	0,8674	0,7461	

Ces résultats sont plutôt étonnants. Pour le pin rouge, le classement selon la corrélation décroissante améliore toutes les mesures de précision. Dans chaque groupe, le classement selon la valeur de x_k croissante rapproche sensiblement Var_{SUN1} de Var_{RHC} et, lorsque $x = d^2h$, $Var_{SUN1} < Var_{RHC}$. Le classement, dans chaque groupe, selon la valeur de x_k décroissante donne des résultats encore meilleurs. Par contraste, le classement des pins ponderosa selon la corrélation décroissante et la valeur de x_k croissante détériore les mesures de précision. Le chiffre exorbitant de 60,861 est attribuable au passage prématuré à un EAS puisque dans ce cas particulier, la valeur de k^* n'est que de 28, ce qui implique que seulement 20% de la population est soumise à un échantillonnage selon un plan π_{pi} . En outre, le classement des pins ponderosa selon la corrélation décroissante et la valeur de x_k décroissante améliore Var_{SUN1} seulement si $x = d^2h$.

Ces résultats indiquent qu'il existe probablement un mode de classement qui a pour effet de minimiser Var_{SUN1} et qui est susceptible de favoriser une plus grande précision qu'un simple classement selon la valeur de X décroissante. Toutefois, ce mode de classement variera normalement selon l'information supplémentaire, et même un classement intuitivement raisonnable pourra donner des résultats imprévus. On ne peut dire à l'heure actuelle s'il existe un mode de classement qui soit optimal – au sens de la minimisation de $Var(\hat{f}_{SUN1})$ – pour le plan π_{pi} approximatif utilisé dans cette étude. À notre connaissance, aucune méthode optimale n'a encore été présentée à ce propos.

Pour toutes les populations, la supériorité de l'estimation du RM par rapport aux trois plans étudiés (π pt authentique, non-PSPT, π pt approximatif) lorsque $X = \text{numéro de séquence de l'arbre est remarquable}$.

Tableau 3
Efficacité relative des plans SUN1, SUN2 et RHC pour les autres populations d'arbres, lorsque l'estimation du rapport de moyennes (RM) sert de point de référence

$n/N\%$	X	n	$\frac{\text{VAR}_{\text{SUN2}}}{\text{EQM}_{\text{RM}}}$	$\frac{\text{VAR}_{\text{SUN1}}}{\text{EQM}_{\text{RM}}}$	$\frac{\text{VAR}_{\text{RHC}}}{\text{EQM}_{\text{RM}}}$
			Pin ponderosa		
1	No	2	1.9608	1.9794	1.9507
1	d^2h	2	0.1050	0.1096	0.1077
2	No	3	2.2976	1.9264	2.2275
2	d^2h	3	0.1768	0.1919	0.1859
5	No	7	2.8717	2.0681	2.7819
5	d^2h	7	0.3113	0.3890	0.3670
10	No	14	3.2528	2.2745	3.0294
10	d^2h	14	0.2928	1.3724	0.4488

Pin rouge ¹					
2	No	2	2.0210	1.9485	2.0029
2	d^2h	2	0.9076	0.9026	0.9104
5	No	5	2.9295	2.3141	2.8236
5	d^2h	5	0.8874	1.3456	0.8991
10	No	9	3.5548	2.0124	3.2958
10	d^2h	9	0.8699	1.3192	0.8942

1	No	5	4.8011	3.7104	4.7625
1	d^2h	5	0.4043	0.4161	0.4174
2	No	9	5.5940	3.7441	5.5044
2	d^2h	9	0.5129	0.5510	0.5476
5	No	22	6.5290	3.3082	6.5253
5	d^2h	22	0.5035	0.6385	0.6085
10	No	44	7.7977	2.6635	6.5708
10	d^2h	44	0.3854	0.7214	0.6146

¹ Le taux de sondage de 1% a été exclus parce qu'on aurait obtenu un échantillon de taille $n = 1$.

On peut voir, d'après la figure 1, que la relation entre le volume du fût et le numéro de l'arbre est tout à fait aléatoire, c.-à-d. que le numéro de séquence ne renseigne aucunement sur le volume du fût. De fait, il y a un incon-venient à utiliser cette information supplémentaire non informative pour déterminer les probabilités de sélection.

Les résultats du tableau 2 ont ceci de curieux qu'ils laissent supposer que SUN2 est moins efficace que RHC ou SUN1 pour certaines combinaisons de variable auxiliaire et de taux de sondage. Peut-être est-ce un effet artificiel du calcul approximatif de certaines probabilités de sélection du second ordre contenues dans VAR_{SUN2} . Il peut aussi s'agir d'un effet du classement utilisé dans SUN1 ou de la taille des groupes utilisés pour l'échantillonnage RHC. On peut calculer la valeur exacte de $\text{Var}(f_{\text{SUN2}}^{\pi})$ pour $n = 2$. Nous l'avons fait pour les populations de pins ponderosa et de pins rouges. Les résultats indiquent que VAR_{SUN2} reflète avec beaucoup de justesse la précision du plan SUN2 mais produit des valeurs un peu prudentes. Nous avons calculé des valeurs se situant entre 0.975 et 0.999 pour le rapport $\text{Var}(f_{\text{SUN2}}^{\pi})/\text{VAR}_{\text{SUN2}}$. Il n'y a aucun moyen d'évaluer l'efficacité de la formule d'approximation VAR_{SUN2} pour des échantillons plus grands.

Comparons maintenant les plans RHC et SUN1 entre eux en nous servant à nouveau du tableau 2. Pour des taux de sondage faibles, VAR_{SUN1} et VAR_{RHC} sont essentiellement équivalents lorsque $X = d^2h$. Cependant, pour des taux plus élevés, avec la même variable auxiliaire, f_{gr} est beaucoup plus efficace dans certains cas, notamment lorsque le taux de sondage est de 10%: dans ce cas, f_{gr} est presque 6 fois plus précis que f_{SUN1} .

Ces résultats nous amènent à penser que plus la relation $x_k \propto y_k$ est forte, plus f_{gr} est précis par rapport à f_{SUN} , à cause surtout de l'effet de k^* sur VAR_{SUN} . De petites valeurs de k^* indiquent un passage hâtif à un EAS et correspondent à de petites valeurs de $\text{VAR}_{\text{SUN2}}/\text{VAR}_{\text{SUN1}}$. En revanche, des valeurs de k^* élevées correspondent à des rapports de variances qui se rapprochent de 1. Pour le tulipier d'Amérique, la méthode SUN1 est telle que seulement les trois quarts de l'échantillon est prélevé selon un plan π pt étant donné un taux de sondage de 10% et $X = d^2h$; nous pensons que le passage prématuré à un EAS explique aussi le faible rendement du plan SUN1 par rapport au plan RHC. Lorsque $X = \text{numéro de séquence de l'arbre}$, SUN1 est beaucoup plus précis que RHC, et sa précision relative s'accroît lorsque n augmente.

La forte amélioration d'efficacité observée lorsque d'autres variables que le numéro de séquence de l'arbre servent de variable auxiliaire est une indication de l'efficacité des méthodes étudiées ici, lorsque X est corrélé positivement avec Y , et de la faiblesse de l'échantillonnage avec probabilité proportionnelle à une variable auxiliaire, lorsque celle-ci n'a aucun rapport avec Y . Les tendances observées pour le tulipier d'Amérique dans le tableau 2 s'observent aussi, de façon générale, dans les résultats concernant les autres essences. Le tableau 3 résume une partie de ces résultats. En ce qui a trait au pin ponderosa, SUN1 est toujours moins précis que RHC lorsque $X = d^2h$, quel que soit le taux de sondage, et SUN2 est toujours supérieur aux deux autres lorsque cette variable est utilisée. Pour toutes les essences, étant donné un taux de sondage de 10% et $X = d^2h$, SUN1 est moins précis que les autres plans et moins efficace que l'estimation du rapport de moyennes, sauf dans le cas des pins à encens.

Tableau 1
Populations d'arbres étudiées dans la comparaison empirique des plans SUN1, SUN2 et RHC
Les quatre dernières colonnes contiennent les coefficients de corrélation de Pearson entre x_k et y_k

Essences	$N^{(1)}$	$t^{(2)}$	d	d^2	d^2h	No
Pin ponderosa	140	9,366.6	0.99	0.99	0.99	0.31
Tulipier d'Amérique <i>Liriodendron tulipifera</i>	336	18,255.5	0.96	0.96	0.99	-0.07
Pin à encens <i>Pinus taeda</i>	437	1,835.8	0.96	0.96	0.99	-0.32
Pin rouge <i>Pinus resinosa</i>	91	4,075.7	0.96	0.96	0.97	-0.05

(1) N = nombre d'arbres qui composent la population.
(2) t = volume total.

3. POPULATIONS D'ARBRES

Le tableau 1 indique les populations d'arbres qui ont été étudiées tandis que la figure 1 décrit la relation entre les diverses modalités de x_k et le caractère étudié pour la population de tulipiers d'Amérique. On voit que le volume a une relation presque parfaitement proportionnelle avec d^2h , nettement curviligne, avec d , et intermédiaire avec d^2 . Pour la base de sondage dont les éléments ne sont pas classés dans un ordre déterminé, aucune tendance parti-culière ne se dessine dans la relation entre le numéro de séquence et le volume. Des observations semblables sont faites pour les trois autres populations d'arbres.

Comme il n'existe aucune observation pour laquelle $nx_k > T_N$ pour les quatre populations et les diverses combinaisons de variable auxiliaire et de taux de sondage, les enregistrements comportent tous une marge d'imprécision.

4. RÉSULTATS

4.1 Comparaison des variances

Le tableau 2 présente une comparaison des variances des estimateurs SUN1, SUN2 et RHC de t pour la population de tulipiers d'Amérique pour chacun des taux de sondage étudiés; pour les besoins de la comparaison, la variance est exprimée en proportion de l'EQM de l'estimateur du RM (rapport de moyennes). Le tableau 3 donne les résultats les plus pertinents pour les trois autres populations d'arbres. En ce qui concerne la méthode SUN1, nous avons classé les éléments des populations par ordre décroissant selon la valeur de X , comme le recommandait Sunter (1977a, 1977b). Examinons d'abord les résultats du tableau 2 relatifs à la population de tulipiers.

Pour un taux de sondage donné, la précision de chaque estimateur étudié par rapport à celle de l'estimateur du RM augmente selon l'ordre de modalités suivant pour $X = No, d, d^2, d^2h, c, a-d$, à mesure que s'accroît le degré de proportionnalité de la variable auxiliaire au volume du fût de l'arbre. Étant donné que l'approximation de la variance de SUN2 donne des résultats satisfaisants, VAR_{SUN2} peut être considérée comme un indice de l'efficacité des plans RHC et SUN1 par rapport à un plan πt authentique. Si le taux de sondage est faible et qu'on

Tableau 2
Efficacité relative des plans SUN1, SUN2 et RHC pour la population de tulipiers d'Amérique, lorsque l'estimation du rapport de moyennes (RM) sert de point de référence

$n/N^{(0)}$	X	n	VAR _{SUN2} EQM _{RM}	VAR _{SUN1} EQM _{RM}	VAR _{RHC} EQM _{RM}	k^*1
-------------	-----	-----	--	--	---	--------

1	No	4	4.8120	3.3136	4.7767	332
1	d	4	0.6735	0.6684	0.6731	333
1	d^2	4	0.4605	0.4596	0.4613	330
1	d^2h	4	0.3361	0.3378	0.3402	330
2	No	7	5.1327	2.6346	5.0568	325
2	d	7	0.7090	0.6982	0.7081	325
2	d^2	7	0.5731	0.5694	0.5751	318
2	d^2h	7	0.4263	0.4542	0.4369	316
5	No	17	5.4938	1.6643	5.2793	309
5	d	17	0.7305	0.7808	0.7283	309
5	d^2	17	0.6541	0.6992	0.6608	291
5	d^2h	17	0.4603	1.2638	0.4935	285
10	No	34	5.8326	1.0985	5.3594	247
10	d	34	0.7385	0.7083	0.7339	247
10	d^2	34	0.6712	0.9687	0.6864	260
10	d^2h	34	0.4298	3.0140	0.5037	250

k^* désigne l'observation, dans la base de sondage ordonnée, à partir de laquelle le plan SUN1 passe d'un échantillonnage avec πt à un EAS.

nous avons défini N_i de telle sorte que certains groupes aient pour taille $N_i = \lfloor N/n \rfloor^{g_{if}}$, où g_{if} désigne la fonction du plus grand entier, et les autres, $N_i = \lfloor N/n \rfloor^{g_{if}} + 1$. On choisit des groupes de chaque catégorie de taille de manière que la somme de la taille des groupes soit égale à N . Si N/n est un entier, tous les groupes sont évidemment de même taille. Dans la suite du texte, nous désignerons l'expression (5) par VAR_{RHC} .

Le plan RHC n'est pas un plan π pt authentique puisque la division de la population introduit un élément de hasard qui n'a aucun rapport avec la valeur de la variable auxiliaire et que l'équation (4) n'est pas un estimateur de Horvitz-Thompson. La probabilité de sélection dépend à la fois de la valeur de X_{ik} et de la probabilité qu'un élément soit inclus dans le groupe i . Le classement des éléments de la population n'a aucun effet sur VAR_{RHC} .

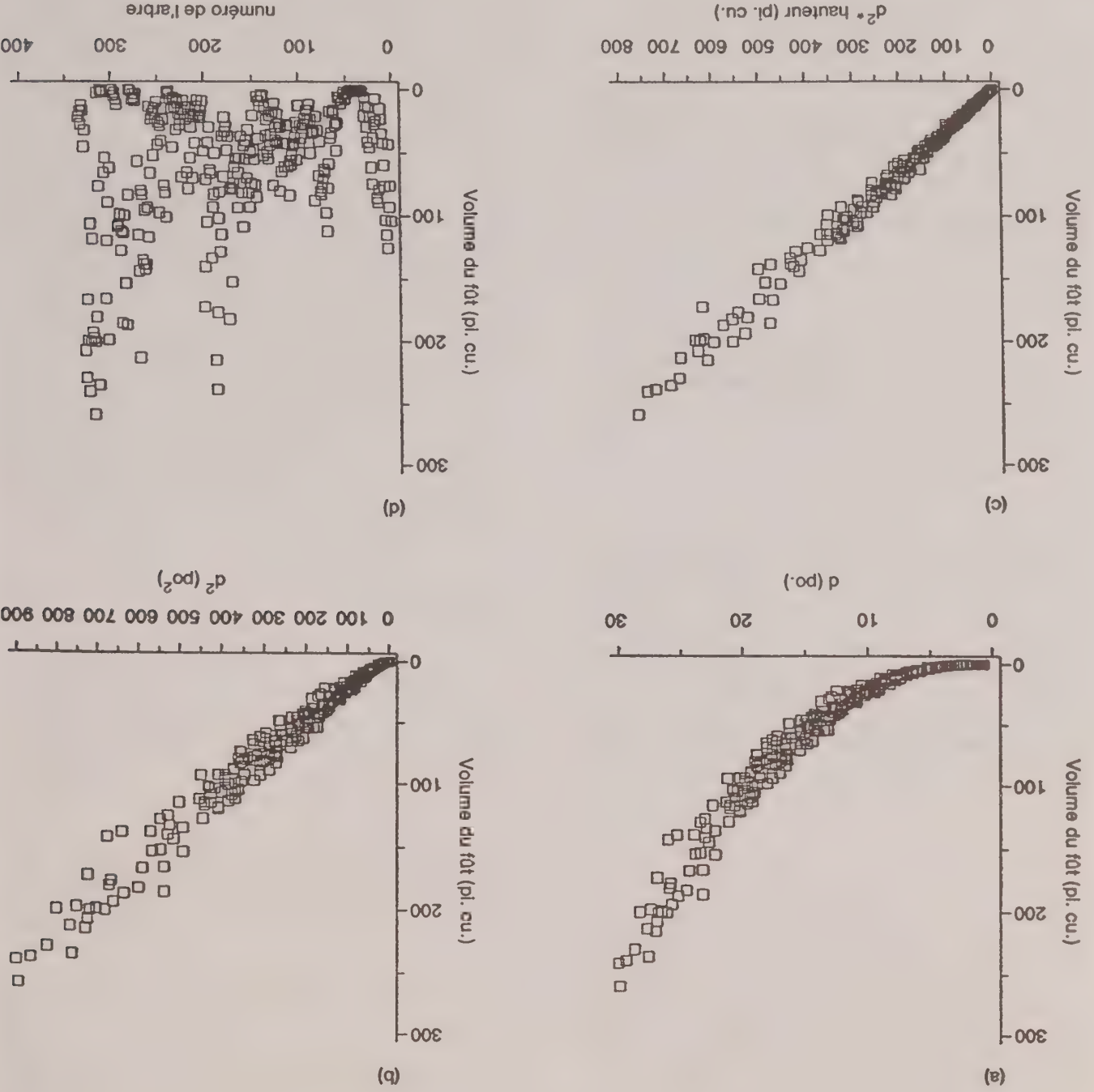


Figure 1. Relation entre le volume du fût et certaines de ses dimensions pour le tulipier d'Amérique: (a) diamètre à hauteur d'homme; (b) diamètre au carré; (c) diamètre au carré fois la hauteur; (d) numéro de séquence de l'arbre.

dans le haut de la liste des éléments de la population. Appliquer la règle de Sunter dans ces circonstances équivalait essentiellement à tirer un échantillon aléatoire simple.

Le π -estimateur du total de population peut être calculé par la formule

$$(1) \quad \hat{f}_{\pi \text{SUN1}} = \sum_{N} \frac{y_k}{Y_k} \pi_k I_k$$

où I_k est la fonction indicatrice d'inclusion dans l'échantillon. La variance est calculée par la formule

$$\text{Var}(\hat{f}_{\pi \text{SUN1}}) = -\frac{1}{2} \sum_{N} \sum_{N}^{k=1} \text{Cov}(I_k, I_l) \left(\frac{\pi_k}{Y_k} - \frac{\pi_l}{Y_l} \right)^2, (2)$$

qui est la formule de Yates-Grundy, où $\text{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$ (Särndal et coll. 1992). Dans le reste du texte, nous désignerons l'expression (2) par VAR_{SUN1} .

2.2 Variante 2 de Sunter

Dans Sunter (1986, 1989), on décrit un plan π pt authentique pour des échantillons de taille $n > 2$. Pour plus de clarté, posons $z_k = x_k / T_N$ et classons les éléments de la population de telle sorte que

$$nz_k < z_k, k = 1, \dots, N - (n + 1)$$

$$(n - k)z_l < z_k, l \geq k \geq N - n,$$

où $z_k = \sum_{N=k}^{N-1} z_l$. Soit m_k le nombre d'échantillons, sur n , qui restent à prélever lorsqu'on arrive au k -ième élément de population u_k . Compte tenu de ce que les deux conditions sont satisfaites, l'algorithme suivant permet de tirer un échantillon π pt authentique. Pour u_k , $P(u_k | m_k) = nz_k / Z_k$ jusqu'à ce que $m_k = 0$ ou que $m_k = N - k$; dans ce dernier cas, on élimine une des unités restantes avec une probabilité de $1 - (m_k z_l / Z_k)$ et on conserve les autres.

Il n'est pas toujours possible de classer les éléments de la population de telle manière que les conditions ci-dessus soient satisfaites. Sunter (1986) décrit un algorithme qui permet de vérifier si un tel classement peut être fait. Les probabilités de sélection sont

$$(3) \quad \begin{aligned} \pi_k &= nz_k \\ \pi_{kl} &= n(n - 1)z_k z_l \gamma_k, k \leq N - n - 1, l > k, \end{aligned}$$

où

$$\gamma_k = \frac{1}{Z_k} \left(1 - \frac{z_1}{Z_1} \right) \dots \left(1 - \frac{z_{k-1}}{Z_{k-1}} \right),$$

$$k = 2, \dots, N - (n + 1).$$

où I_{ik} est la fonction indicatrice d'inclusion dans l'échantillon pour l'élément u_k du groupe i . Le total pour la population est donc estimé au moyen de l'équation

$$(4) \quad \hat{f}_{gr} = \sum_{n}^{l=1} \hat{f}_{i\pi},$$

et la variance de cet estimateur est définie par l'équation

$$\text{Var}(\hat{f}_{gr}) = \frac{1}{N(N-1)} \left(\sum_{n}^{l=1} N_l^2 - N \right)$$

$$(5) \quad \left(\sum_{N}^{k=1} T_N y_{ik}^2 / X_k - t^2 \right).$$

Il convient de souligner que l'équation (5) dépend de la taille des groupes et que sa valeur est minimum lorsque tous les groupes ont la même taille. Dans notre application,

Pour ce qui est des autres probabilités de sélection du second ordre, c'est-à-dire π_{kl} pour $l > k > N - n$, on ne peut les déterminer que par une énumération des échantillons possibles, chose sans doute irréalisable. Sunter prétend qu'on peut obtenir une bonne approximation de ces probabilités avec l'équation (3); de fait, nous sommes servis de cette équation ici. Avec ces probabilités de sélection, $\hat{f}_{\pi \text{SUN2}}$ est exprimé par le membre de droite de l'équation (1). On obtient une approximation de $\text{Var}(\hat{f}_{\pi \text{SUN2}})$ au moyen de l'équation (2), tandis que π_{kl} pour $l > k > N - n$, est calculée à l'aide de l'équation (3). Il convient de souligner les différences entre SUN1 et SUN2 . Dans le cas de SUN1 , les probabilités de sélection composées sont calculées avec exactitude pour toutes les paires, mais, vu l'existence d'un EAS pour certains éléments, il ne s'agit pas d'un plan π pt authentique. En ce qui a trait à la variante 2 de Sunter, il s'agit d'un plan π pt authentique, mais $\text{Var}(\hat{f}_{\pi \text{SUN2}})$ ne peut être calculée qu'approximativement. Nous désignerons cette variante par VAR_{SUN2} .

2.3 Plan RHC

La description du plan RHC est simple; les propriétés de l'estimateur RHC sont largement illustrées dans Rao, Hartley et Cochran (1962) et dans Rao (1966, 1978). Après avoir déterminé la taille d'échantillon n , on divise aléatoirement la population de taille N en n groupes de taille N_i , où $N = \sum_{i=1}^n N_i$ ($i = 1, \dots, n$). Soit X_{ik} l'information supplémentaire relative à l'élément u_k du groupe i , $k = 1, \dots, N_i$, et posons $X_{i.} = \sum_{k=1}^{N_i} X_{ik}$. Un élément est prélevé dans chaque groupe avec une probabilité $p_{ik} = X_{ik} / X_{i.}$. L'estimateur du total pour le groupe i est défini par l'équation

$$\hat{f}_{i\pi} = \sum_{N_i}^{k=1} \frac{y_{ik}}{X_{i.}} I_{ik},$$

proportionnalité des probabilités de sélection π_k pour un sous-ensemble de la population et, dans l'autre cas, il permet une certaine variation de la taille d'échantillon (Sunter 1977a, 1977b; Schreuder et coll. 1990). Afin de conserver le maximum de précision, Sunter suppose, dans le second cas, que la variance de $n(s)$ est faible et, dans le premier cas, que la modification de certaines valeurs π_k n'a pas de conséquences graves. Dans notre étude, nous n'utilisons que la première méthode parce que le plan RHC prévoit lui aussi une taille d'échantillon fixe et que le but de cette étude est précisément de comparer l'applicabilité de la méthode de Sunter à celle de la méthode RHC. Särndal et coll. (1992) décrivent en détail la répartition de l'échantillon et le calcul des probabilités de sélection. Pour une partie de la population, $\pi_k \propto x_k$, où x_k est l'information supplémentaire qui existe sur le k -ième sujet (ou enregistré). Désignons par k^* un élément de la population ordonnée. Alors, pour tous les éléments pour lesquels $k < k^*$, l'échantillonnage est proportionnel à x_k . Le processus se termine lorsque la répartition d'un échantillon de taille n est complétée ou lorsque $k = k^* = \min\{k : nx_k/t_k \geq 1\}$, $N - n + 1$, où $t_k = \sum_{j \geq k} x_j$. Dans ce dernier cas, on tire le reste des échantillons parmi les éléments pour lesquels $k \geq k^*$ selon la méthode d'échantillonnage séquentiel à partir d'une liste de Bebbington (1975). L'avantage de cette méthode, comme le souligne Sunter, est qu'il n'est pas nécessaire de parcourir la base de sondage plus d'une fois. En outre, on peut calculer les probabilités de sélection du premier et du second ordre pendant l'opération de balayage. Comme le plan d'échantillonnage fait que $\pi_{k_l} > 0 \forall k, l; \pi_k \pi_l - \pi_{kl} > 0 \forall k, l$ et que n est fixe, il est facile de calculer l'estimateur de la variance de Yates-Grundy non négatif. Les probabilités de sélection du premier ordre sont $\pi_k = nx_k/T_N$ si $k < k^*$ et $\pi_k = nx_k/T_N$ si $k \geq k^*$, où $T_N = \sum_{k=1}^N x_k$ et $\pi_{k^*} = t_{k^*}/(N - k^* + 1)$. Les expressions pour les probabilités de sélection du second ordre figurent dans Särndal et coll. (1992).

Ainsi, le classement des éléments de la population influe sur l'efficacité du plan SUN1 puisque les probabilités de sélection et, donc, la variance dépendent de k^* (voir (2) ci-dessous). Pour de grands échantillons, la condition $k^* = \min\{k : nx_k/t_k \geq 1\}$, $N - n + 1$ peut devenir $k^* = \min\{k : nx_k/t_k \geq 1\}$, ce qui peut amener la transformation prématurée d'un échantillonnage π pt en un EAS à cause du classement des éléments de la base de sondage. Notons qu'il n'est pas nécessaire que l'inéquation $x_k/t_k < x_{k+1}/t_{k+1}$, pour $k' > k$, soit vraie puisque si $x_k > x_{k+1}$ et si $t_k > t_{k+1}$, le rapport x_k/t_k peut fort bien être plus grand ou plus petit que x_{k+1}/t_{k+1} . Il se peut donc que $nx_k > t_k$ et que $nx_{k'} < t_{k'}$, pour des valeurs k et k' quelconques, où $k' > k$. Dans ce cas, qui peut être assez fréquent, il est difficile de dire si le passage d'un échantillonnage π pt à un EAS devrait se faire dès que $nx_k \geq t_k$ ou non. Il peut arriver que $nx_k \geq t_k$ pour les deux ou trois premiers éléments de la population mais que nx_k soit plus petit que t_k pour la majeure partie de la base de sondage. C'est exactement le cas lorsque n est grand et qu'un petit nombre de valeurs x_k très élevées figurent

Au début, Sunter a proposé deux plans π pt approximatifs; dans un cas, il assouplit la condition touchant la

2.1 Variante 1 de Sunter

2. PLANS D'ÉCHANTILLONNAGE

Tous les plans ont été analysés avec des échantillons de 1, de 2, de 5 et de 10%. L'efficacité des différents plans a été mesurée par la variance de chaque estimateur de $t = \sum_{k=1}^N y_k$. La formule de l'échantillonnage aléatoire simple suivi de l'estimation d'un rapport de moyennes a servi de point de référence puisque cette formule utilise la même information supplémentaire que les autres plans. Nous avons comparé la variance des plans d'échantillonnage décrits dans la section suivante à l'erreur quadratique moyenne de l'estimateur du rapport de moyennes (RM), calculée au moyen de l'approximation du second ordre par la méthode delta décrite dans Sukhatme et coll. (1984).

En foresterie, l'information supplémentaire est souvent un caractère de dimension comme la hauteur (h), le diamètre à hauteur d'homme (d) ou une combinaison des deux; facile à obtenir, cette information peut servir à un échantillonnage efficace en vue d'estimer le volume de fût ou la biomasse, y . Par exemple, la forme d'un tronc d'arbre laisse supposer des rapports entre d , h et le volume total de fût par unité de superficie ou pour un peuplement. Dans notre analyse, le paramètre étudié est le volume du fût de l'arbre qui peuvent être utiles pour l'échantillonnage. En pratique, nous devrions recourir à un échantillonnage à plusieurs degrés, mais, pour les besoins de notre exposé, nous nous limiterons à un échantillonnage à un degré. Pour les plans RHC et SUN, nous nous sommes servis des variables auxiliaires d , d^2 , d^2h et du numéro de séquence de l'arbre. Le numéro de séquence a été utilisé comme variable auxiliaire puisqu'en l'absence d'un classement selon la taille ce numéro n'a évidemment aucun rapport avec le caractère étudié. Il pourrait servir à évaluer l'effet d'une information supplémentaire non informative sur des méthodes diversifiées (voir Rao 1966).

Les méthodes SUN1, SUN2 et RHC conviennent si l'on dispose d'une liste des éléments de la population dont peut être tiré l'échantillon. On ne s'attend pas à avoir une liste complète des valeurs de la caractéristique étudiée y_k , mais on peut faire en sorte que les probabilités de sélection soient proportionnelles à une variable auxiliaire x_k . Autrement dit, connaissant entièrement x_k avant l'échantillonnage, x_k étant, par hypothèse, à peu près proportionnelle à y_k , nous tentons d'obtenir $\pi_k \propto x_k$ tandis que $n = \text{constante}$.

L'objet de cette étude est de comparer empiriquement les échantillons suffisamment grands.

des avantages concrets qu'elles présentent, il serait bon méthodes ne sont plus utilisées; néanmoins, compte tenu de leur efficacité relative. À notre connaissance, ces deux

Solutions de remplacement pour les plans π pt authentiques: une étude comparative

OLIVER SCHABENBERGER et TIMOTHY G. GREGOIRE¹

RÉSUMÉ

L'échantillonnage sans remise à partir d'une liste avec probabilité proportionnelle à une mesure de taille est peu utilisé en foresterie à cause des difficultés d'application de cette méthode, qu'on appelle d'ailleurs "plan π pt" pour nous étudier un plan π pt authentique (variante 2 de Sunter), un plan π pt approximatif (variante 1 de Sunter) ainsi que la méthode des groupes aléatoires de Rao-Hartley-Cochran et nous calculons la variance des estimateurs correspondants du volume total de fût pour quatre populations d'arbres. Les résultats montrent que, par rapport à la méthode de Rao-Hartley-Cochran, la variante 1 de Sunter produit en général des estimations plus précises lorsque la relation entre l'information supplémentaire, x_k , et le caractère étudié, y_k , est *modérée* mais sensible au mode de classement des éléments de la base de sondage, alors que la méthode de Rao-Hartley-Cochran ne nécessite pas de classement particulier pour les éléments de la base de sondage et semble plus efficace que la variante 1 lorsqu'il existe une forte relation linéaire entre x_k et y_k .

MOTS CLÉS: Échantillonnage avec probabilité proportionnelle à la taille; taille d'échantillon fixe; plans π pt approximatifs; comparaison empirique.

1. INTRODUCTION

Rao (1978) classe les méthodes d'échantillonnage avec probabilités inégales sans remise dans deux grandes catégories: (i) celles où les probabilités de sélection π_k sont proportionnelles à la caractéristique étudiée, y_k , et où est utilisé l'estimateur de Horvitz-Thompson \bar{t}_π ; (ii) celles où sont utilisées d'autres statistiques que l'estimateur de Horvitz-Thompson. Les méthodes de la première catégorie sont appelées plans PPT (probabilité de sélection proportionnelle à la taille) et celles de la seconde catégorie, plans non-PPT. Dans des ouvrages récents, par exemple Särndal et coll. (1992), on désigne la probabilité de sélection par p dans le cas de l'échantillonnage avec remise et par π pour l'échantillonnage sans remise. C'est pourquoi, dans cet article, nous appelons les plans d'échantillonnage de la première catégorie "plans π pt authentiques". Les plans PPT et les plans non-PPT ont ceci de commun que dans des conditions de proportionnalité stricte, c.-à-d. $\pi_k \propto y_k$ et $n(s) \equiv \{ \text{constante} \}$, $\text{Var}(\bar{t}) \equiv 0$, où \bar{t} est l'estimateur utilisé dans chaque cas. Pour cette raison, il semble intéressant de tirer un échantillon sans remise lorsque nous intéressons particulièrement à l'utilité que peuvent avoir ces méthodes pour l'échantillonnage en foresterie. Il existe plusieurs plans π pt authentiques; Rao (1978) en fait une description et une analyse détaillée. Cependant, leur mise en application est souvent une tâche complexe, qui entraîne de lourds calculs étant donné la taille des échantillons habituellement étudiés en foresterie. Bon nombre de ces plans π pt authentiques nécessitent une énumération de tous les échantillons possibles ou emploient

des algorithmes qui deviennent de plus en plus complexes à mesure que n augmente. Sampford (1967) décrit un plan simple qui s'applique aux cas où $n \leq 10$. En foresterie cependant, le nombre d'échantillons qui doivent être prélevés à n importe quelle étape d'une enquête est souvent beaucoup plus élevé que cela, même après une stratification. Par conséquent, soit qu'on établisse une version approximative du processus d'échantillonnage π pt de manière à permettre un calcul exact des probabilités de sélection, soit qu'on fasse une approximation des probabilités de sélection un échantillonnage π pt authentique. Rao, Hartley et Cochran (1962) ont décrit un plan non-PPT, aussi appelé méthode des groupes aléatoires, qui a reçu beaucoup d'attention (voir aussi Rao 1966, 1978). Ce n'est pas un plan π pt, puisqu'il utilise un estimateur différent de \bar{t}_π pour que la variance soit nulle lorsque π_k est proportionnel à y_k , mais il est d'une simplicité remarquable. Par ailleurs, la méthode de Sunter (Sunter 1977a, 1977b) est un plan π pt approximatif comme celui évoqué plus haut. Ces deux méthodes seront désignées dans l'analyse qui suit par les sigles RHC et SUN1. Sunter (1986, 1989) a décrit un plan π pt authentique qui peut être appliqué si certaines conditions relatives au classement des éléments de la base de sondage sont respectées et si l'on peut énumérer les échantillons possibles afin de connaître $\pi_{k\ell}$ pour certaines paires d'éléments. On peut éviter l'énumération en utilisant des valeurs approchées de $\pi_{k\ell}$. Cette méthode sera appelée variante 2, ou SUN2, dans notre analyse. Särndal et coll. (1992) disent que les méthodes SUN1 et RHC impliquent une perte d'efficacité comparative à des plans π pt analogues, mais ils ne font pas d'analyse

¹ Oliver Schabenberger et Timothy G. Gregoire, Department of Forestry, Section Forest Biometrics, College of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, U.S.A.

- KRAFT, C.H., LEPAIGE, Y., et VAN EEDEN, C. (1983). Some finite-sample size properties of Rosenblatt density estimates. *La Revue Canadienne de Statistique*, 11, 95-104.
- OH, H.L., et SCHEUREN, F.S. (1983). Weighing adjustments for unit nonresponse. Dans *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin et D.B. Rubin), 2, 143-184. New York: Academic Press.
- RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.
- ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of the density function. *Annals of Mathematical Statistics*, 27, 832-837.
- SÄRNDAAL, C.-E., et HUI, T.-K. (1981). Estimation for non-response situations: to what extent must we rely on models? Dans *Current Topics in Survey Sampling*. (Eds. D. Krewski, R. Platek et J.N.K. Rao), 227-246. New York: Academic Press.
- SÄRNDAAL, C.-E., et SWENSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.
- WEGMAN, E.J. (1972a). Nonparametric probability density estimation: A summary of available methods. *Technometrics*, 14, 533-546.
- WEGMAN, E.J. (1972b). Nonparametric probability density estimation: A comparison of density estimation methods. *Journal of Statistical Computations and Simulations*, 1, 225-245.

les items y_1 et y_2 compte tenu du peu de variation entre les corrélations (0.05). Par contre, on observe l'effet de la corrélation avec la variable auxiliaire sur $V(t)$ et de $B(V(t))$ en comparant les items y_1 et y_3 , puis y_2 et y_3 , les variations entre les corrélations étant plus fortes dans ces deux cas (0.20 et 0.15 respectivement).

En termes d'estimateurs de variance (Tableau 5.2), nous constatons que:

$$V(t_{\text{RegLnp}}^*) > V(t_{\text{Regnp}}^*) > V(t_{\text{Expnp}}^*),$$

comme tel est le cas pour les estimateurs f_{Reg}^* , f_{Regl}^* et f_{Exp}^* . Ce qui est surprenant et bien sûr dû à l'effet des variables auxiliaires sur les composantes de la variance relatives aux mécanismes de réponse, c'est le fait que les estimateurs f_{Expnp}^* surestiment la variance avec des valeurs absolues de $ER(V(t))$ très élevées, alors que les estimateurs par régression f_{Regnp}^* et f_{Reglnp}^* sous-estiment la variance avec des valeurs absolues de $ER(V(t))$ moindres par rapport à celles de f_{Expnp}^* (Tableau 5.3). Pour les estimateurs f_{Expnp}^* , non seulement la variance totale est élevée par rapport à celle des estimateurs par régression mais aussi la contribution relative de la variance d'échantillonnage est basse (Tableau 5.2).

En termes de taux de recouvrement (Tableau 5.4), les estimateurs f_{Expnp}^* donnent des taux observés plus proches des taux théoriques que les estimateurs f_{Regnp}^* et f_{Reglnp}^* . Cependant, les valeurs du biais relatif $BR(t)$ sont plus élevées pour f_{Expnp}^* que pour f_{Regnp}^* et f_{Reglnp}^* , ce qui rend moins fiables les intervalles de confiance.

EN CONCLUSION

(i) Si le but de l'estimation est de réduire le biais et l'erreur quadratique moyenne, tous les estimateurs rajustés pour la non-réponse performant bien par rapport au mécanisme uniforme de réponses (ne rien faire en quelque sorte face à la non-réponse). Le taux de réduction du biais de chaque estimateur par rapport au biais de l'estimateur naïf est d'au moins 66%. Les estimateurs par régression f_{Regnp}^* et f_{Reglnp}^* sont les meilleurs candidats dans l'ensemble des estimateurs considérés (Table 5.1).

(ii) Si le but est de construire des intervalles de confiance, nous avons besoin d'un couple d'estimateurs $[t, V(t)]$ qui minimisent simultanément les biais absolus $|B(t)|$ et $|B(V(t))|$. Les Tableaux 5.1 et 5.2, montrent bien que les estimateurs f_{Regnp}^* et f_{Reglnp}^* sont les plus performants. Ces estimateurs sont en effet moins sensibles à la non-réponse si l'on considère les valeurs de $ER(t)$ et de $ER(V(t))$ (Tableau 5.3). Il reste que le critère de fiabilité des intervalles de confiance $BR(t) < 0.10$ n'est jamais atteint (Tableau 5.4).

(iii) Le comportement des estimateurs rajustés (i) pour l'item y_1 , qui est l'item le plus corréle avec la variable auxiliaire, comparé à l'item y_3 , puis (ii) pour l'item y_2 comparé à l'item y_3 (y_3 étant l'item le moins corréle avec la variable auxiliaire), montre qu'avec de très fortes variables explicatives (pour y_t et pour Θ^{qk}), on peut avoir

BIBLIOGRAPHIE

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

CHICOINEAU, F., PAYEN, J.F., et THÉLOT, C. (1985). Modélisation et redressement des non-réponses: le cas du salaire. *Bulletin de l'Institut International de Statistique*, LI-3, 15, 3, 1-23.

COCHRAN, W.G. (1977). *Sampling Techniques* (3ième éd.). New York: Wiley.

GIOMMI, A. (1985). On the estimation of the individual response probabilities. *Bulletin de l'Institut International de Statistique*, 2, 577-578.

GIOMMI, A. (1987). Méthodes non-paramétriques pour l'estimation des probabilités de réponse individuelles. *Techniques d'enquête*, 13, 137-144.

GROSBRAS, J.-J. (1987b). Les réponses manquantes. Dans *Les sondages*. (Éds. J.-J. Droesbeke, B. Fichet et F. Tassi). Paris: Economica.

JOHNSON, N.L., et KOTZ, S. (1970). *Continuous univariate distributions-I*. New York: Houghton.

KOTZ, P.S. (1987). Nonresponse in a periodic sample survey. *Journal of Business and Economic Statistics*, 5, 287-293.

pouvons également calculer, pour chaque estimateur donné, (v) l'erreur relative $ER(t) [= IB(t)/t]$, (vi) la variance $V(t) [= EQM(t) - (IB(t))^2]$, (vii) le biais relatif $BR(t) [= |IB(t)| / (V(t))^{1/2}]$ ainsi que (viii) l'erreur relative de l'estimation de variance $ER(V(t)) [= IB(V(t)) / V(t)]$ pour examiner la sensibilité des estimateurs de variance à la non-réponse.

5.3 Interprétation des résultats de la simulation globale

I. Les estimateurs prototypes

Les résultats de la simulation confirment la théorie. En effet, pour ces estimateurs, nous constatons ce qui suit, à partir des tableaux (5.1) à (5.4):

- (i) f_{Exp} , f_{Reg} et f_{Reg1} sont approximativement sans biais;
- (ii) $EQM(f_{Reg}) < EQM(f_{Reg1}) < EQM(f_{Exp})$;
- (iii) $V(f_{Reg1}) < V(f_{Reg}) < V(f_{Exp})$ et $IB[V(f_{Reg1})] < IB[V(f_{Reg})] < IB[V(f_{Exp})]$.

Pour ces estimateurs, on s'attendait également à ce que:

- (i) $IEV(f_{Exp}) \approx V(f_{Exp})$, $IEV(f_{Reg}) \approx V(f_{Reg})$ et $IEV(f_{Reg1}) \approx V(f_{Reg1})$;
- (ii) Biais relatif négligeable $[BR(t) < 0.10]$, les taux de recouvrement sont proches des taux théoriques. Les erreurs relatives $ER(t)$ et $ER(V(t))$ sont négligeables, en partie due à la simulation (erreurs dues au nombre limité de répétitions de l'expérience).

Tableau 5.1

Les valeurs de $IB(t)$, $EQM(t)$

	Y_1	Y_2	Y_3
f_{Exp}	0.036	1.690	-0.052
f_{Reg}	-0.020	0.735	-0.019
f_{Reg1}	-0.012	0.319	-0.012
f_{Naif}	-2.037	5.069	-1.937
f_{Expnpx}	-0.690	1.345	-0.777
f_{Expnpr}	-0.601	1.175	-0.709
f_{Regnpr}	0.293	0.785	-0.414
$f_{Reglnpr}$	0.285	0.376	0.407

Tableau 5.2

Les valeurs de $V(t)$, $IB[V(t)]$ et $100*IE[V_1(t)]/IE[V(t)]$

	Y_1	Y_2	Y_3
f_{Exp}	1.689	1.683	29.8
f_{Reg}	0.734	0.697	72.2
f_{Reg1}	0.319	0.293	34.0
f_{Naif}	0.918	0.911	43.3
f_{Expnpx}	0.869	1.403	32.0
f_{Expnpr}	0.814	1.291	35.1
f_{Regnpr}	0.700	0.627	73.9
$f_{Reglnpr}$	0.294	0.259	36.7

Tableau 5.3
Les valeurs de $ER(t)$ et $ER(V(t))$

	Y_1	Y_2	Y_3
f_{Exp}	-0.0024	-0.0015	-0.0040
f_{Reg}	0.0014	-0.0015	-0.0056
f_{Reg1}	-0.0008	-0.0012	-0.0009
f_{Naif}	-0.1377	-0.0083	-0.1474
f_{Expnpx}	-0.0466	0.6141	-0.0591
f_{Expnpr}	-0.0406	0.5860	-0.0540
f_{Regnpr}	-0.0198	0.1038	0.0315
$f_{Reglnpr}$	-0.0193	-0.1191	-0.0310

Tableau 5.4
Les niveaux $P_o(t)$ à 90%, 95% et le $BR(t)$

	Y_1	Y_2	Y_3
f_{Exp}	0.873	0.922	0.027
f_{Reg}	0.881	0.929	0.024
f_{Reg1}	0.866	0.926	0.021
f_{Naif}	0.322	0.427	2.126
f_{Expnpx}	0.851	0.906	0.740
f_{Expnpr}	0.872	0.925	0.666
f_{Regnpr}	0.839	0.908	0.350
$f_{Reglnpr}$	0.804	0.871	0.526

II. L'estimateur naïf

L'estimateur naïf enregistre des valeurs absolues de $IB(t)$ et de $ER(t)$ très élevées par rapport aux autres estimateurs (Tableaux 5.1 et 5.3). Il en est de même des valeurs de $EQM(t)$ (Tableau 5.1). Les valeurs des taux de recouvrement observés $P_o(t)$ aussi bien que celles du biais relatif $BR(t)$ ne sont guère surprenantes compte tenu de la grandeur du biais d'estimation ponctuelle (Tableau 5.4). Le comportement, en termes de variance et d'estimateur de variance (Tableau 5.2) de f_{Naif} provient du fait qu'il constitue un cas particulier de f_{Exp} en supposant des mécanismes de réponse uniforme. Ceci revient en quelque sorte à supposer que les données sont manquantes aléatoirement.

III. Les estimateurs rajustés

La réduction du biais et de l'erreur quadratique moyenne découlant de l'utilisation des estimateurs rajustés (Tableau 5.1) est très importante, en comparaison avec l'estimateur naïf, surtout pour les estimateurs par la régression (les estimateurs f_{Regnpr} et $f_{Reglnpr}$). En termes de variance (Tableau 5.2), nous avons les inégalités suivantes:

$$V(f_{Reglnpr}) < V(f_{Regnpr}) < V(f_{Expnpx}),$$

qui sont analytiquement difficiles à démontrer. On observe peu de variation [en termes de $V(t)$ et de $IB[V(t)]$] entre

Les résultats de cette étude empirique sont illustrés par des schémas de $\mathbb{B}(f_{Expnp}^*)$ et $\mathbb{EQM}(f_{Expnp}^*)$ en fonction de la constante C . Cette étude sommaire nous amène à constater, d'une part que la valeur C_n de la constante C optimale est dans l'intervalle $[0; 1]$, dépend de la taille de l'échantillon et décroît lorsque celle-ci augmente (Figure 5.1 et Figure 5.2).

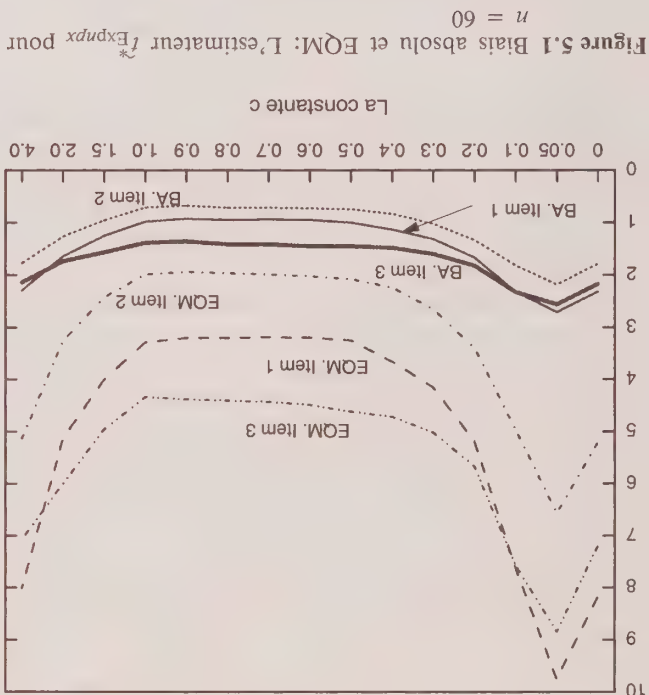


Figure 5.1 Biases absolus et EQM: L'estimateur f_{Expnp}^* pour $n = 60$

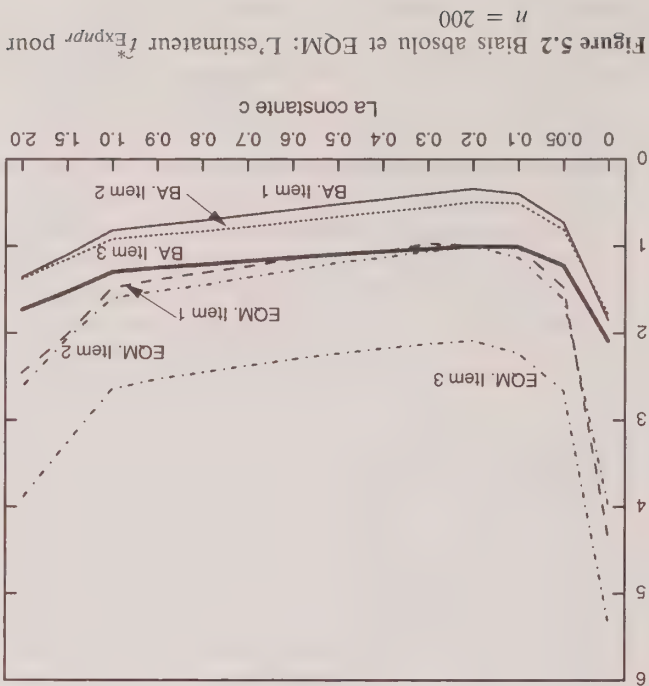


Figure 5.2 Biases absolus et EQM: L'estimateur f_{Expnp}^* pour $n = 200$

Nous constatons également que l'estimateur f_{Expnp}^* est toujours meilleur en termes de moindres biais et erreur quadratique moyenne que l'estimateur f_{Expnp}^* dans l'intervalle $[0; 1]$ tel qu'illustré à titre d'exemple par la figure 5.3 pour l'item 3, l'item le moins corrélé avec la variable auxiliaire. Un fait très important à souligner est que pour l'estimateur f_{Expnp}^* nous atteignons plus rapidement les valeurs du biais et de l'erreur quadratique moyenne de cet intervalle à $C = 4$. Contrairement à l'estimateur f_{Nat}^* dans $[0; 1]$ à $C = 0.05$ et en dehors de valeurs maximales à $C = 0.05$ avant de prendre les valeurs du biais et de l'erreur quadratique moyenne de f_{Nat}^* à $C = 0$. Nous constatons aussi que pour une taille n assez grande et pour toute valeur de C dans l'intervalle $[0; 1]$ la variation est à peine perceptible (Figure 5.3). Pour cette raison, nous suggérons d'utiliser une valeur de compromis: $C = 0.5$ (c'est-à-dire $h = 0.55_{gs}$).

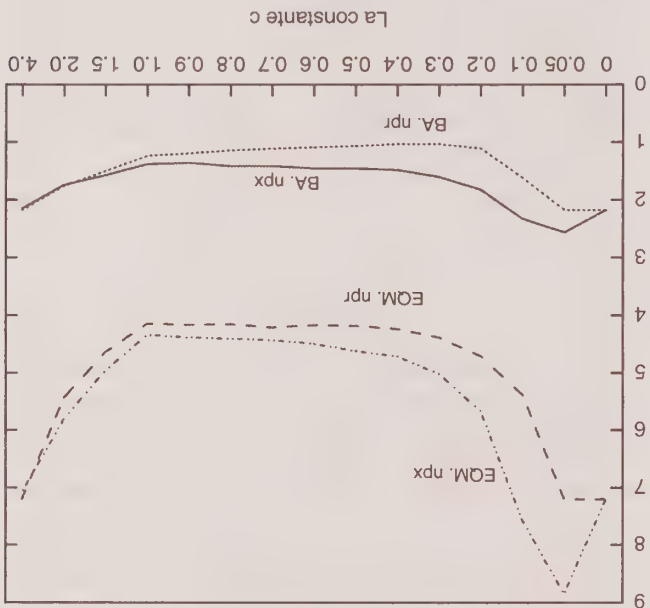


Figure 5.3 Biases absolus et EQM: Les estimateurs f_{Expnp}^* et f_{Expnp}^* pour l'item 3

5.2 Comparaison globale des estimateurs

L'opération complète de la simulation consiste à, (i) tirer d'abord l'échantillon s de taille $n = 200$ de la population de taille $N = 1,000$, (ii) exécuter ensuite les mécanismes de réponse par unité et par item pour obtenir les ensembles r_q ($q = 1, 2, 3$) et (iii) calculer enfin, pour chaque estimateur, les valeurs de t et de $V(t)$. Nous répétons cette opération $1K$ fois. Une fois l'expérience complétée, nous calculons, comme mesures de performance (i) le biais $\mathbb{B}(t) = \mathbb{E}(t) - t_q$, (ii) l'erreur quadratique moyenne $\mathbb{EQM}(t) = \mathbb{E}(t - t_q)^2$, (iii) l'espérance de l'estimateur de variance $\mathbb{E}(V(t))$ et (iv) le taux de recouvrement théorique $P_0(t) = \mathbb{P}\{|t - t_q| \leq Z_{\alpha/2}[V(t)]^{1/2}\}$. Nous

5. ÉTUDE DE MONTE CARLO: COMPARAISON DES ESTIMATEURS

Pour les fins de simulation, nous supposons qu'un schéma de Bernouilli régit chacun des mécanismes de réponse (totale et partielle) et qu'un tirage aléatoire simple sans remise est le plan d'échantillonnage utilisé. Nous considérons un vecteur (y_1, y_2, y_3) de trois items ($Q = 3$) et une variable x contenant l'information auxiliaire. Nous générerons d'abord les $x_k (k \in U)$ par une loi Gamma de paramètres a_1 et a_2 . La génération des items y_1, y_2, y_3 repose sur le modèle linéaire (3.1) et la loi Gamma. Plus précisément, nous générerons les $y_k^q (k \in U \text{ et } q = 1, 2, 3)$ selon une loi Gamma de paramètres $a_1^q(x_k)$ et $a_2^q(x_k)$ définis par:

$$\beta_{2x_k}^{a_2^q} = \frac{a_2^q}{\beta_{2x_k}^{a_2^q}}, \quad a_{2q}(x_k) = \frac{\beta_{2x_k}^{a_2^q}}{a_2^q}, \quad \frac{\beta_{2x_k}^{a_2^q}}{a_2^q}$$

$$a_2^q = \beta_2^2 a_2^q \left\{ \frac{1}{2} \frac{\rho_{xy^q}^2}{\rho_{xy^q}} - 1 \right\}, \quad q = 1, 2, 3.$$

Le choix de la loi Gamma est motivé par sa forme générale dont découle une grande variété de distributions et par le fait qu'elle peut représenter la distribution de plusieurs types de population (Johnson et Kotz 1970, p.172). Nous fixons a priori les paramètres a_1, a_2, β_2^q et ρ_{xy^q} ($q = 1, 2, 3$), à savoir:

$$a_1 = 2, \quad a_2 = 10, \quad (\beta_1, \beta_2, \beta_3)' = (0.75 \ 0.65 \ 0.60)', \quad (\rho_{xy^1} \rho_{xy^2} \rho_{xy^3})' = (0.90 \ 0.85 \ 0.70)'.$$

Pour générer les probabilités de réponse par unité et par item, nous considérons les formes exponentielles suivantes:

$$\phi_k = \exp\{- (\lambda_1 x_k + \lambda_2 v_k)\} \quad \text{et} \quad \psi^{qk} = \exp\{- (\lambda_1^q x_k + \lambda_2^q v^{qk})\},$$

où les v_k et v^{qk} sont issues d'une loi uniforme (0 ; 1). Les constantes $\lambda_1, \lambda_2, \lambda_1^q$ et λ_2^q sont telles que: $\lambda_1 = 0.15/x_U$, $\lambda_1^q = 0.15/\beta_2^q x_U$ et $\lambda_2 = \lambda_2^q = 0.45$ ($q = 1, 2, 3$). Une telle paramétrisation permet d'avoir un taux de réponse (total et partiel) moyen d'environ 70%. L'on aurait pu faire varier ces constantes ou utiliser d'autres fonctions continues.

5.1 Comparaison des deux variantes de l'approche d'ENP

Nous considérons une population de taille $N = 100$ et tirons un échantillon s de taille $n = 60$, auquel nous faisons subir les mécanismes de réponse. Nous répétons l'échantillonnage IK fois et calculons le biais $B(f_{ENP}^*)$ et l'erreur quadratique moyenne $EQM(f_{ENP}^*)$, pour différents valeurs de C ($C \geq 0$). Dans un second temps, nous refaisons cette expérience avec $N = 1,000$ et $n = 200$.

où $h(n) = h(g_k, k \in s)$ est une constante positive qui converge vers zéro à un taux bien approprié. La constante optimale théorique, selon le critère de moindre erreur quadratique moyenne est donnée par $h(n) = K_f n^{-1/5}$ où K_f , telle que définie par Rosenblatt (1956) et Wegman (1972a et b), est obtenue par l'expression $K_f = [9f(x)/2 |f''(x)|^2]^{1/5}$. En pratique, $h(n)$ ne peut s'obtenir que par simulation car elle dépend de la fonction de densité à estimer. Gijmami (1985) a utilisé $h(n) = 2EI_s n^{-1/3}$ où EI_s est l'étendue interquartile dans l'échantillon. Krafé, Lepage et van Eeden (1983) ont pris comme choix, $h(n) = C(n)EI_s$ où $C(n) = (K_f/EI_s)n^{-1/5}$. Nous allons adopter, comme choix, $h(n) = C(n)S_{g_s}$, où $C(n) = (K_f/S_{g_s})n^{-1/5}$ et où S_{g_s} est l'écart type corrigé des valeurs $g_k (k \in s)$. En nous inspirant de l'étude de Krafé, Lepage et van Eeden (1983), nous allons déterminer de façon empirique une valeur C_n de C qui est optimale selon le critère de moindre biais et moindre erreur quadratique moyenne de l'estimateur f_{ENP}^* et comparer les deux versions de l'approche d'ENP.

4.3 Estimateurs par expansion et par la régression

Le calcul du biais et de la variance approximatifs des estimateurs f_{Exp}^{reg} et f_{Reg}^{reg} est simplifié par le fait que les probabilités ϕ_k et ψ^{qk} sont supposées connues. Pour les estimateurs f_{Expnp}^*, f_{Regnp}^* et f_{Reginp}^* , ces probabilités sont estimées par $\hat{\phi}_k$ et $\hat{\psi}^{qk}$. Ces estimateurs de probabilités ne répondent à aucun modèle de probabilité pouvant nous permettre de calculer le biais et la variance conditionnellement à ce modèle. Autrement dit, les ensembles r_q sont générés par des mécanismes de réponse inconnus dont on estime les probabilités de réponse par une approche ne permettant pas l'inférence conditionnellement à un modèle quelconque sous-jacent à l'estimation des probabilités. On serait tenté de recourir au développement en série de Taylor de la fonction $1/\Theta^{qk}$ pour justifier l'approximation de $1/\Theta^{qk}$ par $1/\Theta^{qk}$. Dans ce cas, le biais et la variance de f_{Expnp}^* et f_{Regnp}^* seraient approchés par le biais et la variance approximatifs de f_{Expnp}^*, f_{Regnp}^* et f_{Reginp}^* . Cependant, pour des tailles d'échantillon pas assez grandes, on risque d'avoir $1/\Theta^{qk} \neq 1/\Theta^{qk}$ pour la majorité des $k \in r_q$, et par conséquent:

$$V(f_{Expnp}^*) \neq V(f_{Exp}), \quad V(f_{Regnp}^*) \neq V(f_{Reg}), \quad \text{et} \quad V(f_{Reginp}^*) \neq V(f_{Regl}).$$

Cependant pour construire des intervalles de confiance basés sur f_{Expnp}^*, f_{Regnp}^* et f_{Reginp}^* , il faut définir des estimateurs pour leurs variances respectives. N'ayant pas d'expressions explicites pour ces variances, il est difficile de définir des estimateurs de variance et d'étudier analytiquement leurs propriétés. Le choix de tel ou tel autre estimateur est très délicat à justifier. Le moyen le plus naturel de se doter d'estimateurs de variance pour les variances de f_{Expnp}^*, f_{Regnp}^* consiste à faire une simple substitution de $\Theta^{qk} (= \phi_k \psi^{qk})$, par $\hat{\Theta}^{qk} (= \hat{\phi}_k \hat{\psi}^{qk})$, $\forall k \in r_q$, et de Θ^{qkt} , par $\hat{\Theta}^{qkt}$, $\forall k \neq t \in r_q$ ($\hat{\Theta}^{qkt} = \hat{\phi}_k \hat{\psi}^{qkt}$), dans toutes les formules de variance définies pour les estimateurs de variance respectifs des estimateurs f_{Expnp}^*, f_{Regnp}^* et f_{Reginp}^* .

4. ESTIMATEURS AVEC DES PROBABILITÉS DE RÉPONSE ESTIMÉES

En pratique, les probabilités de réponse φ_k et ψ_{qk} ainsi que les produits de probabilités $\Theta_{qk} = \varphi_k \psi_{qk}$ ($k \in U, q = 1, \dots, Q$) sont plutôt des paramètres à estimer. Nous les estimons par $\hat{\varphi}_k, \hat{\psi}_{qk}$ et $\hat{\Theta}_{qk} = \hat{\varphi}_k \hat{\psi}_{qk}$ respectivement. Nous définissons des estimateurs ayant la même forme que les estimateurs prototypes $f_{\text{exp}}^{\text{reg}}$, f_{reg} et $f_{\text{reg}}^{\text{reg}}$ vus dans la section 3 en ayant soin de remplacer les paramètres inconnus par leurs estimations respectives. Nous dénotons ces estimateurs par f_{expnp}^* , f_{regnp}^* et f_{reginp}^* respectivement. Les estimateurs de variance découlent des expressions (3.3), (3.4) et (3.6) dans lesquelles les paramètres inconnus sont substitués par leurs estimations.

4.1 Estimation des probabilités de réponse

Les probabilités φ_k et ψ_{qk} sont, en théorie, des fonctions des variables auxiliaires, c'est-à-dire des fonctions de la forme $\varphi_k = f_1(v, z_k)$ et $\psi_{qk} = f_2(\mu_q, x^{qk})$ où les quantités v et μ_q ($q = 1, \dots, Q$) sont des paramètres inconnus et où le couple de vecteurs (z, x^q) , c'est-à-dire, $[(z_1, x_1^q), \dots, (z_k, x_k^q), \dots, (z_N, x_N^q)]$, contient l'information auxiliaire disponible pour chaque item y^q . L'approche d'estimation non-paramétrique utilise uniquement l'information contenue dans (z, x^q) pour estimer les φ_k et ψ_{qk} . Nous considérons ici le cas particulier où les $z_k = x^{qk} = x_k$, $\forall q$ ($q = 1, \dots, Q$), et $\forall k \in s$. Soit $x_s = \{x_k : k \in s\}$ l'ensemble d'information auxiliaire relatif à l'échantillon. Définissons $\tau_s = \{\tau_k : k \in s\}$ un ensemble de fonctions telles que $\tau_k : \mathbb{R}^n \rightarrow \mathbb{R}^1$, pour tout k dans s . Notons par $g_k = \tau_k(x_s)$, $\forall k \in s$, la valeur de la k -ième fonction évaluée en x_s . Subdivisons s en n groupes s_k pas nécessairement disjoints et dont les tailles respectives sont données par :

$$n_k = \sum_{j \in r_q} D(g_k - g_j), (k \in s), \quad \left\{ \begin{array}{ll} 1 & \text{si } |g_k - g_j| \leq h_k, \\ 0 & \text{sinon,} \end{array} \right.$$

pour une constante h_k donnée qui peut dépendre de toutes les valeurs g_k ($k \in s$). L'ensemble $s_k = \{j : g_j \in [g_k \pm h_k]\}$, $\forall k \in s$, contient les unités j dont les valeurs g_j varient très peu entre elles. On appelle ce groupe, le groupe dont l'unité k est le noyau ou simplement le k -ième groupe. Autrement dit, s_k est un sous-ensemble de s pour lequel les valeurs de x tombent dans le voisinage de $x = x_k$ au sens de la distance euclidienne que définit $d(k, j) = |\tau_k(x_s) - \tau_j(x_s)| \leq h_k = h(g_k)$, c'est-à-dire que $s_k = \{j : d(k, j) \leq h_k\}$. Posons $r_k = s_k \cap r$ et $r_{qk} = s_k \cap r^q$. Les effectifs respectifs de ces ensembles sont m_k et m_{qk} où :

$$m_k = \sum_{j \in r} D(g_k - g_j), (k \in r); \quad m_{qk} = \sum_{j \in r_q} D(g_k - g_j), (k \in r_q, q = 1, \dots, Q).$$

$$\hat{\varphi}_k = \frac{m_k}{m_k}, \forall k \in r; \quad \hat{\psi}_{qk} = \frac{m_{qk}}{m_{qk}}, \forall k \in r_q, \quad (4.1)$$

alors que le produit $\Theta_{qk} = \varphi_k \psi_{qk}$ est estimé par le taux :

$$\hat{\Theta}_{qk} = \hat{\varphi}_k \hat{\psi}_{qk} = m_{qk}/n_k, (k \in r_q, q = 1, \dots, Q), \quad (4.2)$$

qui n'est rien d'autre que le taux de réponse dans le k -ième groupe. Cette simplification du produit estimé $\hat{\Theta}_{qk} = \hat{\varphi}_k \hat{\psi}_{qk}$ n'est cependant possible que lorsque les deux mécanismes de réponse sont régis par les mêmes variables auxiliaires.

Deux approches sont considérées ici, celle basée sur les valeurs de la variable x (np) et celle basée sur les rangs des valeurs de la variable x (npr). L'ENP(npx), proposée par Giommi (1987), s'obtient en prenant $g_k = \tau_k(x_s) = x_k$ ($k \in s$). Pour pallier à un effet éventuel des trop grandes et des trop petites valeurs de x_s , nous introduisons une variante qui consiste à utiliser les rangs de x_s , c'est-à-dire l'ENP(npr). Considérons la fonction u telle que $u(z) = 1$ si $z \geq 0$ et $u(z) = 0$ si $z < 0$. Pour toute unité k dans s , posons $u_k = \sum_s u(x_k - x_j) =$ le nombre de composantes de x_s qui sont inférieures ou égales à $x_k =$ le rang de x_k dans s . L'ENP(npr) est alors équivalente à poser $g_k = \tau_k(x_s) = u_k$ ($k \in s$).

4.2 Choix des limites des intervalles

Le problème majeur dans l'approche d'ENP est le choix optimal des constantes h_k qui déterminent les limites des intervalles $[g_k - h_k; g_k + h_k]$, $\forall k \in s$, c'est-à-dire, un choix de $h_k = h(g_s)$ qui réduit le biais et l'erreur quadratique moyenne de tout estimateur utilisant les produits estimés $\hat{\Theta}_{qk}$ définis par la formule (4.2). Selon Giommi (1985, 1987), les termes n_k, m_k et m_{qk} dont on se sert pour estimer les probabilités de réponse sont, mis à part les facteurs de normalisation, des estimateurs par la méthode du noyau de la fonction de densité selon l'approche de Rosenblatt (1956) pour diverses séries de valeurs de g . À titre d'exemple, il est facile de montrer que :

$$n_k = \sum_{j \in s} D(g_k - g_j) = 2nh(n)f_n(g_k),$$

$\Theta_k^q = \mathbb{P}(k \in r_q | s)$ alors, (i) $\pi_k^q \neq \pi_k \Theta_k^q$ et (ii) $\pi_k \Theta_k^q \neq \pi_k \Theta_k^q$. De plus, (iii) $\Theta_k^q = \Theta_k^q$ si les probabilités ψ_{qk} sont indépendantes de r et (iv) $\pi_k^q = \pi_k \Theta_k^q$ si les φ_k ne dépendent pas de s et si les ψ_{qk} ne dépendent ni de r ni de s .

3. QUELQUES ESTIMATEURS SPÉCIAUX

Supposons qu'il existe une variable auxiliaire x_q (pour le q -ième item) fortement corrélée avec la variable y_q et telle que x_q^k est connue $\forall k \in s$ ou $\forall k \in U$. Prenons le cas particulier où $x_q^k = x_k$, $\forall q (q = 1, \dots, Q)$ et posons le modèle linéaire ξ suivant

$$(3.1) \quad \begin{cases} \mathbb{E}_\xi(y_{qk} | x_k) = \beta^q x_k \\ \text{Cov}_\xi(y_{qk}, y_{q\ell} | x_k, x_\ell) = \begin{cases} 0 & \text{si } k = \ell \\ \sigma_2^q x_k & \text{sinon} \end{cases} \end{cases}$$

Swenson (1987).

Résultat 1. Si x_k est connue, $\forall k \in s$, alors l'estimateur par la régression, dénoté par \hat{r}_{reg} et défini par:

$$(3.2) \quad \hat{r}_{\text{reg}} = \left(\Sigma \frac{y_{qk} \pi_k \Theta_{qk}^q}{x_k} \right) / \left(\Sigma \frac{\pi_k \Theta_{qk}^q}{x_k} \right) \Sigma \frac{\pi_k}{x_k},$$

est approximativement sans biais pour t_q . Sa variance approximative est une somme de trois composantes V_1 , V_2 et V_3 représentant les parts respectives de la variance dues aux phases de sélection, c'est-à-dire:

$$V_2 = \mathbb{E} \left\{ \Sigma \Sigma \Delta^{\varphi_{k\ell}} (E^{qk} / \pi_k \varphi_k) (E^{q\ell} / \pi_\ell \varphi_\ell) \right\},$$

$$V_3 = \mathbb{E} \mathbb{E} \left[\Sigma \Sigma \Delta^{\psi_{qk\ell}} (E^{qk} / \pi_k \Theta_{qk}^q) (E^{q\ell} / \pi_\ell \Theta_{q\ell}^q) | s \right],$$

où les E^{qk} sont les résidus théoriques du modèle (3.1). Un estimateur de $V(\hat{r}_{\text{reg}})$ est donné par $V(\hat{r}_{\text{reg}}) = V_1 + V_2 + V_3$ (ou $V_2^+ = V_2 + V_3$) avec:

$$(3.3) \quad V_1 = \Sigma \Sigma \Delta^{\pi_{k\ell}} \frac{r_q \pi_{k\ell} \Theta_{qk\ell}^q}{y_{qk}^q} \left(\frac{\pi_k}{y_{qk}^q} \right) \left(\frac{\pi_\ell}{y_{q\ell}^q} \right),$$

$$(3.4) \quad V_2^+ = \Sigma \Sigma \Delta^{\Theta_{qk\ell}^q} \frac{r_q \Theta_{qk\ell}^q}{e_{qk}^q} \left(\frac{\pi_k \Theta_{qk}^q}{e_{qk}^q} \right) \left(\frac{\pi_\ell \Theta_{q\ell}^q}{e_{q\ell}^q} \right),$$

et

$$(3.5) \quad \hat{r}_{\text{reg1}} = N X_U \left(\Sigma \frac{y_{qk} \pi_k \Theta_{qk}^q}{x_k} \right) / \left(\Sigma \frac{\pi_k \Theta_{qk}^q}{x_k} \right),$$

est approximativement sans biais pour t_q . Sa variance approximative est aussi une somme de trois composantes V_1 , V_2 et V_3 . L'expression de $V_1(\hat{r}_{\text{reg1}})$ diffère de celle de $V_1(\hat{r}_{\text{reg}})$ par l'utilisation des résidus théoriques E_{qk}^q à la place des valeurs brutes y_{qk} , alors que les expressions de V_2 et V_3 sont identiques à celles définies ci-haut pour \hat{r}_{reg} . Un estimateur de $V(\hat{r}_{\text{reg1}})$ est donné par $V(\hat{r}_{\text{reg1}}) = V_1 + V_2^+$ où:

$$(3.6) \quad V_1 = \Sigma \Sigma \Delta^{\pi_{k\ell}} \frac{r_q \pi_{k\ell} \Theta_{qk\ell}^q}{e_{qk}^q} \left(\frac{\pi_k}{e_{qk}^q} \right) \left(\frac{\pi_\ell}{e_{q\ell}^q} \right),$$

et où $V_2^+ = V_2 + V_3$ est obtenue par la formule (3.4).

Remarque 1. Si $x_k = 1$, $\forall k \in U$, la formule (3.5) définit un estimateur, dénoté par \hat{r}_{Exp} où:

$$(3.7) \quad \hat{r}_{\text{Exp}} = N \Sigma \frac{y_{qk} \pi_k \Theta_{qk}^q}{y_{qk}^q} / \Sigma \frac{r_q \pi_k \Theta_{qk}^q}{y_{qk}^q} = \frac{N}{N} \Sigma \frac{r_q \pi_k \Theta_{qk}^q}{y_{qk}^q}.$$

L'estimateur \hat{r}_{Exp} est appelé "estimateur par expansion". Un estimateur de variance approximativement sans biais pour $V(\hat{r}_{\text{Exp}})$ découle des formules (3.4) et (3.6).

Remarque 2. Si on prend $\Theta_{qk} = \Theta_{qk}^q (0 < \Theta_{qk}^q \leq 1)$, $\forall k \in U$, dans la formule (3.7), on obtient un estimateur, dénoté par \hat{r}_{Naif} , appelé "estimateur naïf". Son expression est donnée par:

$$(3.8) \quad \hat{r}_{\text{Naif}} = N \Sigma \frac{r_q \pi_k}{y_{qk}^q} / \Sigma \frac{r_q \pi_k}{1}.$$

Si les π_k sont constants, l'expression (3.8) devient identique à la formule (3.5) dans laquelle on pose $\Theta_{qk} = \Theta_{qk}^q (0 < \Theta_{qk}^q \leq 1)$, $\forall k \in U$, et $x_k = 1$, $\forall k \in U$.

Remarque 3. Pour les quatre estimateurs définis ci-haut, les modèles sous-jacents découlent du modèle général (3.1) et sont les suivants: $y_{qk} = \beta^q x_k + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ et $V(\epsilon_{qk}) = \sigma_2^q x_k$ pour les deux premiers, $y_{qk} = \beta^q + \epsilon_{qk}$, $\mathbb{E}(\epsilon_{qk}) = 0$ et $V(\epsilon_{qk}) = \sigma_2^q$ et N connue pour les deux derniers. Pour l'estimateur naïf, il faut ajouter le modèle uniforme de réponses par unité et par item.

Estimation non-paramétrique des probabilités de réponse en théorie de l'échantillonnage

THÉOPHILE NIVONSENGA¹

RÉSUMÉ

Nous traitons le problème de non-réponse en s'inspirant du modèle de sélection en phases proposé par Särndal et Swenson (1987). Pour estimer les probabilités de réponse, nous recourons à l'approche d'estimation non-paramétrique initiée par Giommi (1987). Nous définissons des estimateurs sous le modèle d'estimation non-paramétrique et étudions empiriquement leurs propriétés générales. L'inférence est basée sur la notion de quasi-randomisation (Oh et Scheuren 1983). L'accent est mis sur l'estimation de la variance et la construction des intervalles de confiance. Nous constatons, par une étude de Monte Carlo, qu'il est possible d'améliorer la qualité des estimateurs considérés en utilisant une variante de l'approche d'ENP. Elle permet également de confirmer la performance des estimateurs par régression en termes d'estimation de la variance.

MOTS CLÉS: Pondération par phases; estimateurs par régression; estimateurs de variance.

1. INTRODUCTION

Pour contrer les effets de la non-réponse sur l'estimation de paramètres d'une population finie, nous considérons le phénomène de la non-réponse comme un processus de sélection des unités en trois phases. Nous recourons, par conséquent, à la pondération par phases. Cette procédure d'ajustement affecte à chaque unité observée, un poids inversement proportionnel à la probabilité de figurer dans l'échantillon, à la probabilité de réponse par unité étant donné l'échantillon et à la probabilité de réponse par item étant donné l'échantillon et l'ensemble des répondants par unité. En pratique, seules les probabilités d'inclusion dans l'échantillon sont connues. Le problème auquel on est confronté, est celui d'estimer les probabilités individuelles de réponse avant de les incorporer dans les formules des estimateurs d'intérêt. L'approche d'estimation non-paramétrique est l'une des procédures d'estimation des probabilités de réponse. Elle est motivée par l'utilisation de variables auxiliaires liées aux mécanismes de réponse par unité et par item (Giommi 1985, 1987) et pouvant être corrélées avec les variables d'intérêt. L'on évite ainsi de supposer que la non-réponse est indépendante des variables à l'étude (Oh et Scheuren 1983). Cette approche permet également d'éviter de postuler un ou des modèles paramétriques régissant la réponse tels que les modèles Logit et Tobit (Grosbras 1987b; Chicoineau, Payen et Thélot 1985) ou les modèles de réponses homogènes par groupe (Oh et Scheuren 1983; Särndal et Swenson 1985, 1987). Nous considérons, dans l'étude de Monte Carlo illustrant certains estimateurs sous l'approche d'estimation non-paramétrique, le cas très particulier où les deux mécanismes de réponse sont régis par les mêmes variables auxiliaires. La différence entre les items va résider dans le degré de corrélation entre chaque item et les variables auxiliaires.

2. LA NON-RÉPONSE: UN PROCESSUS DE SÉLECTION EN TROIS PHASES

Soit $U = \{1, 2, \dots, k, \dots, N\}$ une population finie de taille N . Soit s un échantillon de taille fixe n tiré de U par un plan $\phi(s)$ connu et caractérisé par les probabilités d'inclusion $\pi_k > 0$, $\forall k$ et $\pi_{kl} > 0 \forall k \neq l$. On veut observer les unités $k \in s$ par rapport à un ensemble de \bar{Q} items $y_1, \dots, y_q, \dots, y_{\bar{Q}}$ ($\bar{Q} \leq 1$) puis estimer le total par item $t_q = \sum U y_{qk}$, pour tout $q (\bar{q} = 1, \dots, \bar{Q})$. Supposons que, conditionnellement à s , chaque unité k participe à l'enquête avec probabilité $\phi_k > 0$ et que la probabilité que deux unités k et l participent est $\phi_{kl} > 0$ avec $\phi_{kk} = \phi_k$. Dénotons par r l'ensemble des unités qui acceptent de participer à l'enquête et par $\phi(r | s)$ le mécanisme ayant permis d'obtenir l'ensemble r . Supposons également que, conditionnellement à s et r , chaque unité $k \in r$ répond à l'item y_q avec probabilité $\psi_{qk} > 0$ et que la probabilité que deux unités k et $l \in r$ répondent à l'item y_q soit $\psi_{qkl} > 0$ avec $\psi_{qkk} = \psi_{qk}$. Dénotons par r_q l'ensemble des unités qui, ayant accepté de participer à l'enquête, répondent à l'item y_q et par $\phi(r_q | s, r)$ le mécanisme par lequel on obtient l'ensemble r_q pour tout $q (q = 1, \dots, \bar{Q})$. Les ensembles s , r et r_q sont issus des trois phases de sélection pour lesquelles seules les probabilités d'inclusion dans s sont connues. La composition des mécanismes de sélection des unités donne lieu aux produits de probabilités que nous dénotons par $\pi_k \theta_{qk}$ où $\theta_{qk} = \phi_k \psi_{qk}$ et $\theta_{qkl} = \phi_{kl} \psi_{qkl}$ avec $\theta_{qkk} = \theta_{qk}$, qui ne correspondent pas à des probabilités d'inclusion. La quantité θ_{qk} ne correspond pas non plus à une probabilité d'inclusion pour les deux phases de réponse conditionnellement à s . En effet, si on définit les probabilités d'inclusion dans r_q par $\pi_{r_q}^* = \mathbb{P}(k \in r_q)$ et les probabilités d'inclusion dans r_q par $\pi_{r_q}^* = \mathbb{P}(k \in r_q)$

¹ Théophile Nivonsenga, Ph.D., Chercheur, Centre de Recherche Clinique, Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, PQ, Canada, J1H 5N4.

5. CONCLUSIONS

Du point de vue du biais, l'estimateur par quotient de stratification a posteriori (POS) est essentiellement sans biais dans presque toutes les petites régions. De plus, l'estimateur dépendant de la taille de l'échantillon (DT) a des valeurs négligeables pour le biais dans presque toutes les petites régions. Les estimateurs synthétique (SYN) et composite (COM) ont des valeurs de biais beaucoup plus élevées que les autres estimateurs.

Du point de vue de l'efficacité, les estimateurs SYN et COM ont invariablement des RCEQM sensiblement plus faibles que les autres estimateurs. L'estimateur DT est beaucoup plus efficace que l'estimateur POS et a de plus, pour quatre des quatorze régions, des valeurs de la RCEQM qui sont proches de celles des estimateurs SYN et COM. En outre, quand on considère l'estimateur COM, le problème du calcul du α optimum se pose. En pratique, on ne peut utiliser qu'une valeur estimée de α , ce qui entraîne une baisse de l'efficacité de cet estimateur. Donc, si l'on tient compte à la fois du biais et de l'efficacité, il semblerait que l'estimateur DT soit préférable aux autres estimateurs étudiés dans le contexte de l'EPA au Frioul. Les taux d'échantillonnage au Frioul sont relativement élevés et l'importance de l'efficacité et du biais relatifs de ces estimateurs peut être différente dans des régions où les taux d'échantillon-nage sont faibles, p. ex. le Piémont et la Lombardie.

BIBLIOGRAPHIE

DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada. *Techniques d'enquête*, 8, 19-52.

FABBRIS, L., RUSSO, A., et SANETTI, I. (1988). Storia e proposte in tema di campionamento a livello regionale, provinciale e sub-provinciale per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 4. Dipartimento di Scienze Statistiche, Università di Padova.

FALORSI, P.D., et RUSSO, A. (1987). Un metodo di stima sintetica per piccoli territori dominati indagini ISTAT sulle famiglie. *Atti del Convegno della Società Italiana di Statistica*, Perugia, Italia, 11-20.

FALORSI, P.D., et RUSSO, A. (1989). Un'analisi comparativa di alcune tecniche di stima per piccole aree per l'indagine sulle forze di lavoro. *Rapporto di ricerca FOLA*, 18. Dipartimento di Scienze Statistiche, Università di Padova.

FALORSI, P.D., et RUSSO, A. (1990). La stima dell'errore quadratico medio di alcune forme di stimatore sintetico nei campioni a due stadi utilizzati nelle indagini ISTAT sulle famiglie. *Giomate di studio: Classificazione ed analisi dei dati, metodi, software, applicazioni*, Pescara, Italia, 27-39.

FALORSI, P.D., et RUSSO, A. (1991). Evaluation of small area estimation techniques for Italian Labour Force Survey. *Proceedings of the 1991 Annual Research Conference*. U.S. Bureau of the Census, 80-106.

GHANGURDE, P.D., et SINGH, M.P. (1978). Evaluation of efficiency of synthetic estimates. *Proceedings of the Social Statistics Section, American Statistical Association*, 52-61.

GONZALEZ, M.E., et HOZA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

LEVY, P.S. (1979). Small area estimation synthetic and other procedures, 1968-1978. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 4-19.

PURCELL, N.J., et LINACRE, S. (1976). Techniques for the Estimation of Small Area Characteristics. Document présenté au 3rd Australian Statistical Conference, Melbourne.

SÄRNDA, C.-E., et HIDIRGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SCHAI, W.L. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.

SCHAI, W.L. (1979). A composite estimator for small area statistics. *Synthetic Estimates for Small Areas*, National Institute on Drug Abuse, Research Monograph, No. 24, U.S. Government Printing Office, Washington, D.C., 36-83.

B. Mesures du rendement par petite région

Les tableaux 2 et 3 présentent les biais relatifs en pour-

centage (p ARB) ainsi que la racine carrée de l'erreur qua-

dratique moyenne en pourcentage (p RCEQM) des

estimateurs pour chacune des quatorze régions de services

de santé du Frioul. De plus, le tableau 2 donne le rapport

en pourcentage entre la population de la RSS et la popu-

lation de l'ensemble H des strates qui comprennent la RSS

et la population de la région du Frioul (p_3). Les conclu-

sions qui suivent ressortent de l'examen de ces tableaux.

i) Les estimateurs SYN et COM sont très biaisés dans

certaines petites régions, c'est-à-dire dans celles où le

modèle qui permet d'obtenir l'estimateur SYN est mal

adapté. On trouve généralement un biais considérable

pour les petites régions ayant de faibles valeurs du

rapport p_1 (p. ex. les RSS 1, 2, 3, 4 et 6). Inversement,

des valeurs élevées du rapport p_1 sont associées à de

faibles valeurs pour le biais (p. ex. les RSS 5, 9, 10 et

13). Toutefois, les estimateurs SYN et COM ont inva-

riablement des RCEQM assez faibles pour être intéres-

santes comparativement à celles des autres estimateurs.

Dans trois des quatorze régions (c.-à-d. dans les

régions 3, 4 et 8), l'estimateur COM est invariablement

l'estimateur le plus efficace. Dans deux régions (les

RSS 10 et 12), l'estimateur SYN est manifestement

plus efficace et, dans les régions qui restent, les deux

estimateurs sont à peu près semblables du point de vue

de l'efficacité. De plus, nous observons que les valeurs

les plus faibles de la RCEQM pour l'estimateur SYN

sont en général associées aux valeurs les plus élevées du

rapport p_3 (p. ex. les RSS 1, 2, 5, 6, 9 et 13). Bien que

pour les RSS 3 et 4 on trouve des valeurs élevées pour

le rapport p_3 , on trouve aussi une valeur élevée pour

la RCEQM. Cela est attribuable au biais considérable.

ii) Dans le cas de l'estimateur POS, on trouve des valeurs

négligeables pour le biais dans presque toutes les

petites régions. Les valeurs de la RCEQM pour l'esti-

mateur POS sont beaucoup plus élevées que celles des

autres estimateurs dans toutes les petites régions. Nous

observons une corrélation négative entre la RCEQM

de l'estimateur POS et le rapport p_2 . Cette corrélation

négative est attribuable au fait que la taille de l'échan-

tilion prévue augmente à mesure qu'augmente le

rapport p_2 . Par conséquent, la variance (principale

composante de l'EQM de l'estimateur POS) diminue.

iii) L'estimateur DT a un biais négligeable dans sept des

quatorze petites régions (les RSS 5, 7, 9, 10, 11, 12

et 13). Dans les autres régions, le biais est assez faible.

De plus, dans neuf régions (les RSS 2, 3, 4, 5, 9, 10,

11, 12 et 13), l'estimateur DT a un biais semblable à

celui de l'estimateur POS. L'estimateur DT est préfé-

rable à l'estimateur POS, du point de vue de l'EQM.

Dans quatre régions (les RSS 7, 8, 9 et 13), la RCEQM

est semblable à celle des estimateurs SYN et COM.

iv) Enfin, nous remarquons que dans les régions les plus

grandes, qui ont les valeurs les plus élevées pour le

rapport p_2 (p. ex. les RSS 9 et 5), tous les estimateurs

études donnent des résultats semblables pour ce qui

est du biais et de l'EQM. Pour les autres régions, où

les estimateurs ont des rendements différents, le choix

du meilleur estimateur pose un problème.

Tableau 2

Biais relatif en pourcentage (p ARB) de chacune des quatorze régions de services de santé (RSS) du Frioul pour les chômeurs, selon l'estimateur

Estimateur	DT	COM	SYN	POS	p_1	RSS
	-3.01	-7.68	-10.92	-1.57	19.1	1
	-4.79	-6.97	-9.21	-5.61	16.1	2
	5.79	17.98	28.82	-5.21	15.3	3
	2.99	15.02	20.92	-2.50	16.3	4
	-0.28	0.98	1.61	-0.46	47.1	5
	-3.28	-9.06	-12.24	-1.37	24.6	6
	-1.66	-3.40	-6.25	0.05	81.8	7
	2.17	6.63	11.80	0.81	70.7	8
	0.78	0.68	0.76	0.47	92.2	9
	-1.02	0.51	-1.34	0.36	71.2	10
	-1.62	-5.00	-5.64	-1.01	21.7	11
	-1.19	-6.05	-6.66	-1.22	40.6	12
	-1.28	-1.11	-3.12	-0.95	56.3	13
	-3.53	-3.03	-6.21	-2.51	21.8	14

p_1 = rapport en pourcentage entre la population de la RSS et la population de l'ensemble H des strates qui comprennent la RSS.

Tableau 3

Racine carrée de l'erreur quadratique moyenne en pourcentage (p RCEQM) de chacune des quatorze régions des services de santé (RSS) du Frioul pour les chômeurs, selon l'estimateur

Estimateur	DT	COM	SYN	POS	p_3	RSS
	32.39	21.12	20.41	52.23	19.9	1
	38.30	20.81	19.45	63.36	19.2	2
	42.46	30.71	36.57	57.44	23.2	3
	36.88	27.02	30.09	58.19	23.2	4
	17.87	14.01	13.38	18.81	42.9	5
	22.69	17.00	17.49	28.09	34.8	6
	22.67	21.67	21.47	23.83	6.9	7
	27.40	26.35	28.54	28.75	4.8	8
	16.89	16.40	16.15	17.29	22.9	9
	59.27	53.31	50.12	67.00	1.8	10
	30.42	19.20	18.35	49.82	3.2	11
	33.18	24.04	22.10	46.40	10.7	12
	17.88	15.40	15.53	20.13	22.4	13
	36.81	22.94	23.58	57.80	10.1	14

p_2 = rapport en pourcentage entre la population de la RSS et la population de la région du Frioul.

p_3 = rapport en pourcentage entre la population de l'ensemble H des strates qui comprennent la RSS et la population de la région du Frioul.

La constante F est choisie de façon à faire ressortir l'apport de la composante synthétique. Plus la valeur de F est élevée, moins la partie synthétique est importante. Le choix de la valeur de F dépendrait de plusieurs facteurs. Dans notre étude, nous avons examiné l'efficacité de l'estimateur dépendant de l'échantillon pour $F = 1$. Cette valeur s'est révélée efficace tout en offrant une protection contre le biais de l'estimateur synthétique.

Le raisonnement qui sous-tend l'estimateur DT est le fait que lorsque la taille de l'échantillon dans le domaine d et le groupe g est petite, alors la valeur estimée directe-ment pour le domaine d et le groupe g serait instable, et une estimation synthétique pourrait donner un meilleur résultat. Toutefois, si l'échantillon dans le domaine d et le groupe g est plus grand que prévu, cela ne pose pas de difficulté, puisque le rendement de la partie avec stratification a posteriori directe s'améliorerait à mesure qu'augmenterait la taille de l'échantillon. En conclusion, nous observons que l'estimateur DT peut être considéré comme une forme particulière de l'estimateur par régression dépendant de la taille de l'échantillon donné dans Sarnadal et Hidiroglou (1989), qui possède de bonnes propriétés conditionnelles.

4. DESCRIPTION DE L'ÉTUDE EMPIRIQUE

4.1 Simulation du plan d'échantillonnage de l'EPA

Dans notre étude, nous avons considéré les 14 RSS de la région du Frioül comme de petites régions. La variable qui nous intéresse, y , est le nombre de chômeurs.

Pour évaluer le rendement des divers estimateurs dont nous avons parlé dans la section 3, nous avons eu recours à un plan d'échantillonnage (échantillonnage à deux degrés avec stratification des UPE) identique à celui qui a été adopté pour l'EPA dans le Frioül. Ce plan d'échantillonnage est basé sur le choix de 39 UPE et de 2,290 UPE tirées d'une population de 219 UPE et de 465,000 UPE.

Nous avons choisi indépendamment 400 échantillons répétés de Monte Carlo de taille identique (pour le nombre d'UPE et d'USF) de l'échantillon de l'EPA. Toute l'information utilisée dans la simulation est tirée du recensement général de la population de 1981, de sorte que dY est connue.

4.2 Évaluation des estimateurs pour petites régions

Nous désignons par $dY(mr)$ l'estimation de la dY totale pour la petite région d du r^e échantillon de Monte Carlo quand nous utilisons l'estimateur m . Le biais relatif en pourcentage de l'estimateur m pour la petite région d est donné par

$${}^d\text{ARB}_m = \frac{1}{R} \left(\sum_{r=1}^R {}^dY(mr) - 1 \right) 100,$$

où R est le nombre d'échantillons ($R = 400$). La moyenne du biais relatif absolu en pourcentage de l'estimateur m pour tout l'ensemble de petites régions est:

$$\text{ARB}_m = \frac{1}{D} \sum_{d=1}^D |{}^d\text{ARB}_m|,$$

où D est le nombre de petites régions étudiées ($D = 14$). La racine carrée de l'erreur quadratique moyenne en pourcentage de l'estimateur m pour la petite région d est

$${}^d\text{RCEQM}_m = \frac{{}^d\text{EQM}_m}{100},$$

où l'erreur quadratique moyenne de l'estimateur m pour la petite région d est exprimée par

$${}^d\text{EQM}_m = \frac{1}{R} \sum_{r=1}^R ({}^dY(mr) - {}^dY)^2.$$

La moyenne de la racine carrée de l'erreur quadratique moyenne en pourcentage de l'estimateur m pour toutes les régions est

$$\text{RCEQM}_m = \frac{1}{D} \sum_{d=1}^D {}^d\text{RCEQM}_m.$$

4.3 Analyse des résultats

A. Mesures du rendement global

Les biais absolus moyens en pourcentage ainsi que les moyennes de la racine carrée des erreurs quadratiques moyennes en pourcentage des estimateurs pour petites régions dans le cas de la caractéristique "nombre de chômeurs" de l'EPA sont présentées dans le tableau 1. Les résultats qui suivent ressortent de l'examen de ce tableau.

i) Comme on s'y attendait, l'estimateur POS a le biais le plus faible. Le biais de l'estimateur SYN est plus élevé que celui des autres estimateurs. Le biais de l'estimateur COM est d'environ 30 % inférieur à celui de l'estimateur SYN. Le biais de l'estimateur DT n'est que légèrement inférieur à celui de l'estimateur POS.

ii) C'est pour les estimateurs SYN et COM que les moyennes de la racine carrée des erreurs quadratiques moyennes en pourcentage sont le plus faibles, mais ces estimateurs ont des biais très élevés. L'estimateur POS, avec un faible biais, est, inversement, l'estimateur le moins efficace. La moyenne de la racine carrée de l'erreur quadratique moyenne en pourcentage de l'estimateur DT est d'environ 30 % plus élevée que celle des estimateurs SYN et COM.

Tableau 1

Biais relatif absolu moyen en pourcentage ARB et

moyenne de la racine carrée de l'erreur quadratique moyenne en pourcentage RCEQM pour les chômeurs, selon l'estimateur

Estimateur	ARB	RCEQM
POS	1.75	42.08
SYN	8.97	23.80
COM	6.00	23.57
DT	2.39	31.08

Ici, ${}^dP_g^h$ désigne la population totale pour le groupe d'âge-sexe g dans la petite région d recoupée par la strate h , et δ_{hi} est une variable binaire qui a la valeur 1 si l'UPF hi appartient à la petite région d et la valeur 0 dans le cas contraire. Pour une meilleure explication de la formule (1), nous remarquons que l'UPF est un sous-ensemble d'une petite région et qu'elle ne coupe donc pas cette dernière. L'estimateur par quotient de stratification a posteriori est non biaisé sauf pour l'effet du biais de l'estimation par quotient, qui est habituellement négligeable. L'estimateur est défini comme étant zéro quand il n'y a pas d'échantillon dans le domaine. Cet estimateur n'est pas fiable pour les petites tailles d'échantillon.

3.2 Estimateur synthétique

Pour le calcul d'un estimateur synthétique, on suppose que les moyennes de la population de petites régions pour des sous-groupes donnés de la population sont approximativement égales aux moyennes des populations de régions plus grandes pour les mêmes sous-groupes. Cet estimateur est obtenu au moyen d'une procédure à deux étapes: i) par rapport à un niveau territorial groupé, on détermine des estimations des caractéristiques étudiées pour des sous-groupes de la population; ii) les estimations pour la région au niveau territorial regroupé sont alors modifiées en proportion de la fréquence du sous-groupe dans le petit domaine qui nous intéresse.

L'estimateur synthétique a une faible variance puisque l'est basé sur un échantillon plus grand, mais il comporte un biais qui dépend de la mesure dans laquelle on s'écarte de l'hypothèse d'homogénéité, pour chaque sous-groupe, entre la petite région et la plus grande région par rapport à la caractéristique qui nous intéresse, y . Les problèmes associés aux estimateurs synthétiques ont été relevés par Purcell et Linacre (1976), Gonzalez et Hoza (1978), Changuerde et Singh (1978), Schaible (1979) et Levy (1979), entre autres.

Dans cette étude, nous considérons la forme suivante d'estimateur synthétique (SYN):

$${}^dY_{\text{SYN}} = \sum_{g=1}^G \frac{{}^dP_g}{{}^dP} {}^dP_g, \quad (2)$$

où

$${}^dY_g = \sum_{h=1}^H \sum_{n_h} {}^nY_{gh} K_{nh} {}^nY_{gh}; {}^dP_g = \sum_{h=1}^H \sum_{n_h} {}^nY_{gh} K_{nh} {}^nY_{gh}.$$

3.3 Estimateur composite

L'estimateur composite (COM) étudié ici est obtenu par une combinaison linéaire de l'estimateur SYN (biaisé avec une faible variance de l'échantillon) et de l'estimateur POS (moins biaisé mais avec une variance de l'échantillon élevée):

$${}^dY_{\text{COM}} = \alpha {}^dY_{\text{POS}} + (1 - \alpha) {}^dY_{\text{SYN}}, \quad (3)$$

où α est une constante ($0 \leq \alpha \leq 1$). Cet estimateur minimise la probabilité d'avoir des situations extrêmes (du point de vue tant du biais que de la variance de l'échantillon). Par conséquent, dans une situation concrète donnée, un tel estimateur pourrait se révéler plus avantageux que ses deux composantes considérées séparément.

La valeur optimale de α qui minimise l'EQM de l'estimateur COM est donnée par

$$\alpha_{\text{opt}} = \frac{\text{EQM}({}^dY_{\text{SYN}}) - E({}^dY)({}^dY_{\text{SYN}} - {}^dY)}{\text{EQM}({}^dY_{\text{SYN}}) + \text{EQM}({}^dY_{\text{POS}}) - 2E({}^dY)({}^dY_{\text{SYN}} - {}^dY)}. \quad (4)$$

De plus, quand on ne tient pas compte du terme de covariance dans (4), en supposant que ce terme sera petit par rapport à $\text{EQM}({}^dY_{\text{SYN}})$ et à $\text{EQM}({}^dY_{\text{POS}})$, on peut obtenir de façon approximative le poids optimal α à l'aide de

$$\alpha_{\text{opt}}^* = \frac{\text{EQM}({}^dY_{\text{SYN}})}{\text{EQM}({}^dY_{\text{SYN}}) + \text{EQM}({}^dY_{\text{POS}})}. \quad (5)$$

C'est la méthode utilisée par Schaible (1978) pour définir les poids.

Dans notre travail, les valeurs optimales de α ont été obtenues à partir des données du recensement et à l'aide de la formule (5). Quand nous considérons une véritable enquête par sondage, on ne peut utiliser qu'une valeur estimée du α optimal, ce qui entraîne une diminution de l'efficacité.

3.4 Estimateur dépendant de la taille de l'échantillon

L'estimateur dépendant de la taille de l'échantillon est un cas particulier de l'estimateur composite. La combinaison linéaire de l'estimateur synthétique et de l'estimateur moins biaisé est effectuée pour chaque sous-groupe et dépend du résultat de l'échantillon donné. Nous considérons la forme suivante d'estimateur dépendant de la taille de l'échantillon (DT), qui tient compte de la taille de l'échantillon obtenu dans la petite région. Cet estimateur est défini (Drew, Singh et Choudhry 1982) par

$${}^dY_{\text{DT}} = \sum_{g=1}^G \alpha_g \left(\frac{{}^dP_g}{{}^dP} {}^dP_g \right) + (1 - \alpha_g) \frac{{}^dP_g}{{}^dP} {}^dP_g, \quad (6)$$

où:

$$\alpha_g = \left\{ \begin{array}{l} 1 / ({}^dR_g F) \\ 1 / {}^dR_g > F, \end{array} \right. \quad \text{avec } {}^dR_g = {}^dP_g / {}^dP_g.$$

dans les autres cas

(7)

3. ESTIMATEURS POUR PETITES RÉGIONS

Par rapport à la région géographique générale, nous supposons que la population P est répartie dans D petites régions qui ne se chevauchent pas $1, \dots, d, \dots, D$ pour lesquelles on doit obtenir des estimations. Chaque région est obtenue par regroupement de municipalités. Le problème étudié est l'estimation du total d'une variable y pour toutes les unités qui font partie de la petite région d . En pratique, la petite région d ne recoupera qu'un certain nombre de strates du plan de sondage que nous désignerons par $H = \{h \mid {}^dP_h > 0\}$, où dP_h représente la partie de P_h qui appartient à la petite région d .

Désignant par dN_h le nombre d'UPÉ qui appartiennent à la petite région d de la strate h , nous cherchons à estimer le total pour la petite région

$${}^dY = \sum_{G=1}^g \sum_{h=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} {}^gY_{ghij}.$$

L'élaboration d'une méthode d'estimation particulière pour de petites régions dépend fondamentalement des renseignements disponibles. En Italie, il y a très peu de renseignements disponibles au niveau des petites régions. Actuellement, les renseignements territoriaux disponibles sont les chiffres de population selon le sexe pour chaque municipalité, renseignements recueillis à partir des statistiques administratives. D'ici peu (à la fin de 1994), les chiffres de population par groupe d'âge-sexe seront disponibles pour chaque municipalité. C'est pourquoi, dans cette étude, nous ne considérons que les estimateurs pour petites régions qui utilisent, comme renseignements auxiliaires, le total de la population selon le groupe d'âge-sexe.

3.1 Estimateur par quotient de stratification a posteriori

Un estimateur par quotient de stratification a posteriori (POS) de dY est donné par:

$$({}^dY)^{\text{POS}} = \sum_{G=1}^g \frac{{}^dY_g}{{}^dP_g} {}^dP_g, \quad (1)$$

où

$$({}^dY_g) = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} K_{hij} Y_{ghij} \delta_{hi},$$

$$({}^dP_g) = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} K_{hij} P_{ghij} \delta_{hi},$$

$$({}^dP_g) = \sum_{H=1}^H \sum_{h=1}^{{}^dP_{gh}} = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} P_{ghij},$$

Les USE sont choisies sans remise et avec probabilités égales à partir des UPÉ choisies indépendamment. Tous les membres de chaque ménage dans l'échantillon sont dénombrés.

2.2 Estimateur du total

Par rapport à la région géographique générale, nous introduisons les indices inférieurs suivants: h , pour la strate ($h = 1, \dots, H$); i , pour l'unité primaire d'échantillonnage; j , pour les unités secondaires d'échantillonnage; g , pour les groupes d'âge-sexe ($g = 1, \dots, G$). Dans cette étude, nous considérons les classes d'âge suivantes: 14-19 ans, 20-29 ans, 30-59 ans, 60-64 ans et plus de 65 ans. Une quantité qui se rapporte à l'unité secondaire d'échantillonnage j de l'unité primaire d'échantillonnage i de la strate h sera désignée comme étant la quantité hij et une quantité qui se rapporte à l'unité primaire d'échantillonnage i de la strate h sera désignée comme étant la quantité dans hi .

On utilise aussi les notations suivantes: N_h , pour le nombre d'UPÉ dans h ; P_h , pour le nombre total de personnes dans h ; n_h , pour le nombre d'UPÉ dans l'échantillon choisi dans h ; M_{hi} , pour le nombre d'USE dans hi ; P_{hi} , pour le nombre total de personnes dans hi ; m_{hi} , pour le nombre d'USE dans l'échantillon choisi dans hi ; P_{ghij} , pour le nombre de personnes du groupe g qui appartiennent à hij ; P_{ghij} , pour le nombre de personnes dans hij .

De plus, soit

$$Y = \sum_{G=1}^g \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} Y_{ghij}$$

le total de la caractéristique y pour la population régionale, où X_{ghij} désigne le total de la caractéristique qui nous intéresse, y , pour les P_{ghij} personnes. En fait, l'estimation de Y est obtenue à l'aide d'un estimateur de stratification a posteriori. Cet estimateur est donné par:

$$Y = \sum_{G=1}^g \frac{Y_g}{{}^dP_g} {}^dP_g,$$

où

$$Y_g = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} K_{hij} Y_{ghij}; \quad P_g = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} K_{hij} P_{ghij}$$

représentent des estimations non biaisées de

$$Y_g = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} Y_{ghij}; \quad P_g = \sum_{H=1}^H \sum_{N_h=1}^{{}^dN_h} \sum_{M_{hi}=1}^{{}^dM_{hi}} P_{ghij}.$$

Dans les formules qui précèdent, le poids de base, désigné par le symbole K_{hij} , est exprimé par:

$$K_{hij} = \frac{P_h}{{}^dP_h} \frac{{}^dP_{hi}}{m_{hi}}.$$

Comparaison empirique de méthodes d'estimation pour petites régions pour l'enquête sur la population active italienne

P.D. FALORSI, S. FALORSI et A. RUSSO¹

RÉSUMÉ

L'objet de cette étude était d'évaluer divers estimateurs pour petites régions dans le but de produire des estimations de niveau pour des domaines non planifiés tirés de l'enquête sur la population active italienne. Dans notre étude, les petites régions sont les régions des services de santé, qui sont des domaines territoriaux infrarégionaux non planifiés, qui n'ont pas été isolés au moment de l'établissement du plan d'échantillonnage et qui chevauchent donc les limites des strates du plan d'échantillonnage. Nous considérons les estimateurs suivants: l'estimateur par quotient de stratification a posteriori, l'estimateur composite exprimé sous forme d'une combinaison linéaire de l'estimateur synthétique et de l'estimateur par quotient de stratification a posteriori et l'estimateur moyens en pourcentage et la moyenne des erreurs quadratiques moyennes relatives ont été obtenus par une étude de Monte Carlo dans laquelle le plan d'échantillonnage a été simulé à l'aide de données provenant du recensement de l'Italie de 1981.

MOTS CLÉS: Estimateurs pour petites régions; domaines non planifiés; biais; erreur quadratique moyenne (EQM); étude de simulation.

1. INTRODUCTION

En Italie, comme dans beaucoup d'autres pays, il y a un besoin croissant de données courantes et fiables sur les petites régions. Ce besoin de renseignements concerne la plupart des enquêtes par sondage réalisées par l'Institut national de statistique italien (ISTAT), particulièrement l'enquête sur la population active (EPA), qui a été étudiée pour assurer l'exactitude des estimations régionales.

Dans le passé, la solution adoptée par l'ISTAT pour résoudre ce problème consistait à élargir l'échantillon sans modifier la méthode d'estimation (Fabbri et coll. 1988). Ces dernières années toutefois, pour trouver une solution aux aspects négatifs des échantillons trop gros, on a fait des recherches dans le but d'élaborer des méthodes d'estimation pouvant améliorer l'exactitude des estimations pour petites régions (Falorsi et Russo 1987, 1989, 1990 et 1991).

Dans notre étude, les petites régions sont les régions des services de santé (RSS), qui sont des domaines territoriaux infrarégionaux non planifiés, qui n'ont pas été isolés au moment de l'établissement du plan d'échantillonnage et qui chevauchent donc les limites des strates du plan d'échantillonnage. La taille de ces domaines territoriaux est telle que la fiabilité des estimations ordinaires aurait souffert si ces domaines avaient été conçus avec des tailles d'échantillon fixes distinctes à partir de domaines particuliers.

L'étude a été faite pour évaluer d'autres estimateurs pour petites régions qui peuvent être utilisés pour produire des estimations, au niveau des RSS, à partir de l'EPA. Nous considérons les estimateurs suivants: l'estimateur par quotient de stratification a posteriori, l'estimateur composite exprimé sous forme

d'une combinaison linéaire de l'estimateur synthétique et de l'estimateur par quotient de stratification a posteriori) et l'estimateur dépendant de la taille de l'échantillon. Pour tous les estimateurs considérés ici, les biais relatifs moyens en pourcentage et la moyenne des erreurs quadratiques moyennes (EQM) relatives ont été obtenus par une étude de Monte Carlo dans laquelle le plan d'échantillonnage de l'EPA a été simulé à l'aide de données provenant du recensement de l'Italie de 1981.

2. BRÈVE DESCRIPTION DE LA STRATÉGIE

UTILISÉE POUR CHOISIR L'ÉCHANTILLON DE L'EPA

2.1 Conception

L'EPA est basée sur un plan d'échantillonnage à deux degrés stratifié en fonction des unités primaires d'échantillonnage (UPP). Les UPP sont les municipalités, alors que les unités secondaires d'échantillonnage (USB) sont les ménages. Dans le cadre de chaque région géographique, les UPP sont réparties selon les provinces. Dans chaque province, les UPP sont réparties en deux principaux types de régions: la région autoréprésentative, composée des UPP les plus grandes, et la région non autoréprésentative, composée des UPP les plus petites.

Toutes les UPP de la région autoréprésentative sont échantillonnées, alors que dans la région non autoréprésentative le choix des UPP est effectué dans les strates qui ont à peu près la même taille. Deux UPP de l'échantillon sont choisies dans chaque strate sans remise et avec probabilité proportionnelle à la taille (le nombre total de personnes).

¹ P.D. Falorsi, chercheur principal, Institut national de statistique, Rome, Italie; S. Falorsi, chercheur, Institut national de statistique, Rome, Italie; Aldo Russo, professeur adjoint, Université de Molise, Campobasso, Italie.

REMERCIEMENTS

L'auteur tient à souligner le nom de George T. Duncan, professeur à l'Université Carnegie Mellon, à qui on doit la notion de masque de matrice et qui a contribué à la réalisation d'une version antérieure de cet article, ainsi que celui de Sumittra Mukherjee, étudiant de doctorat de Duncan, qui a fait une lecture critique de l'article et qui a élaboré quelques-unes des formulations qui y sont présentées. Les recherches préliminaires qui ont été faites sur le sujet ont été rendues possibles en partie grâce à une subvention (SES 91-10512) de la National Science Foundation. Les idées exprimées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement les principes directeurs ou les pratiques de la United States Environmental Protection Agency.

BIBLIOGRAPHIE

- BETHLEHEM, J.G., KELLER, W.J., et PANNKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.
- COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 398, 520-524.
- COX, L. (1991). Comment (sur Duncan, G.T. et R.W. Pearson 1991), *Statistical Science*, 6, 232-234.

- COX, L., et ERNST, L. (1982). Controlled rounding. *INFOR*, 20, 423-432.
- DALENIUS, T., et REISS, S. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73-85.
- DUNCAN, G.T. (1990). Inferential disclosure-limited microdata dissemination. *Proceedings of the Survey Research Section, American Statistical Association*, 440-445.
- DUNCAN, G.T., et LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7, 207-217.
- DUNCAN, G.T., et PEARSON, R.W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, 6, 219-239.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1994). Report on disclosure limitation methodology. Statistical Policy Working Paper 22, Office of Management and Budget, Washington, DC.
- STRUDLER, M., OH, L., et SCHEUREN, F. (1986). Protection of taxpayer confidentiality with respect to the tax model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 375-381.
- TENDICK, P. (1991). Optimal noise addition for preserving confidentiality in multivariate data. *Journal of Statistical Planning and Inference*, 27, 341-353.

au multiple de 5 supérieur) défait la concordance qui doit exister normalement entre un total et la somme de ses éléments, et on peut donc préférer l'arrondissement contrôlé, conçu pour préserver cette concordance dans les tableaux à une ou à deux dimensions (Cox et Ernst 1982). Il existe aussi des méthodes pour l'arrondissement contrôlé non biaisé dans ces deux types de tableaux (Cox 1987).

La perturbation de données contribue à la protection du secret statistique en modifiant légèrement les valeurs de microdonnées. La perturbation additive consiste à augmenter une valeur initiale par l'addition d'une valeur de perturbation. Les valeurs de perturbation sont souvent tirées aléatoirement d'une distribution qui a une moyenne nulle et une variance faible par rapport à celle des données. On a recours aussi à la perturbation non aléatoire.

L'arrondissement et la perturbation additive peuvent être assimilés à des masques de décalage. Pour chaque valeur x_{ij} , on calcule le facteur de décalage c_{ij} selon l'algorithme d'arrondissement ou l'algorithme de perturbation, avec $c_{ij} = 0$ pour les valeurs qui n'ont pas besoin d'être modifiées. Alors, $X' = X + C$ est la matrice des valeurs arrondies (ou perturbées).

3.5 Topcodage d'attributs

Étant donné une valeur (élevée) T_j de l'attribut j préalable, le "topcodage" d'attributs est une méthode qui consiste à remplacer toutes les valeurs $x_{ij} > T_j$ par T_j . Étant donné $x_{ij} = f_{ij}T_j + r_{ij}$, f_{ij} étant le quotient entier et r_{ij} , le reste, $0 \leq r_{ij} < T_j$, on calcule $t_{ij} = (\text{Max}\{r_{ij}, (T_j + 1)f_{ij} - 1\}) \bmod(T_j + 1)$. Pour "topcoder" X , on utilise le masque de décalage $\text{Tco}(X) = (t_{ij} - x_{ij})$.

4. REPRÉSENTATION DE MASQUES DE DONNÉES

COMME DES MASQUES DE MATRICE

4.1 Sélection et modification de combinaisons

d'attributs et d'enregistrements

Les formulations exposées dans la section précédente, fondées sur des masques de matrice élémentaires, s'appliquent à tout le fichier de microdonnées X et ne permettent pas un masquage sélectif de sous-ensembles arbitraires d'enregistrements (lignes) ou d'attributs (colonnes) de X . Il est important de pouvoir traiter sélectivement des valeurs de microdonnées dans des sous-ensembles de X (c.-à-d. appliquer sélectivement des masques de données à des sous-matrices de X) afin de préserver le caractère confidentiel des données. Cela peut se faire par une combinaison de masques de matrice élémentaires qui permet de sélectionner des sous-ensembles de lignes et de colonnes dans X , les masques de matrice élémentaires étant définis comme dans les sections précédentes. L'opération se déroule en trois étapes.

À la première étape, on applique le "masque d'annulation" $\text{Ign}(\bar{Q}, R) = AXB$, où A est la matrice de dimension $n \times n$, $A = \sum_{i \in \bar{Q}} U_{ii}$ et B , la matrice de dimension $p \times p$, $B = \sum_{j \in R} U_{jj}$. La matrice A laisse intactes les valeurs contenues dans les lignes de X qui ont été sélectionnées et

5. CONCLUSIONS

Nous avons exposé une approche fondée sur l'algèbre matricielle pour formuler les principales méthodes de protection du caractère confidentiel des microdonnées. Les questions touchant le calcul (par ex., dans le cas de fichiers volumineux) n'ont pas été abordées. Cependant, les méthodes de partitionnement exposées dans la section 4.1 peuvent servir à réduire le volume effectif de calculs lorsqu'on travaille avec des fichiers très volumineux. Le masquage de matrice offre une structure complète à l'intérieur de laquelle les organismes statistiques peuvent développer, évaluer et appliquer des logiciels de protection du caractère confidentiel des microdonnées qui soient fiables. Les organismes pourraient d'ailleurs se partager ces logiciels. Aux États-Unis, un groupe d'experts encouragera les organismes statistiques américains à trouver de nouvelles formes d'application pour les masques de matrice (Federal Committee on Statistical Methodology 1994, p. 82). L'usage généralisé des masques de matrice aura pour conséquence de normaliser les méthodes dont disposent les organismes pour préserver le caractère confidentiel des microdonnées et d'accroître, pour chaque organisme, les possibilités d'évaluation et d'application de ces méthodes.

4.3 Permutation de données

La permutation de données est une méthode qui consiste à permuter certaines valeurs entre des ensembles déterminés d'enregistrements de sorte que certains tableaux à une ou à plusieurs dimensions demeurent inchangés (Dalenius et Reiss 1982). Si on pose $M = \text{Reo}(P)$, où la règle de permutation est donnée par une permutation P des enregistrements touchés, on trouve dans la section 4.1 un masque de matrice pour l'opération de brouillage de Strudler, Oh et Scheuren (1986).

4.2 Brouillage

Lorsque l'opération M consiste en une mise en micro-moyenne, la formulation de la section 4.1 offre un masque de matrice pour l'opération de brouillage de Strudler, Oh et Scheuren (1986).

$$\tilde{X} = M(\text{Ign}(\bar{Q}, R)) + X - \text{Ign}(\bar{Q}, R).$$

avaient été "annulées" par l'opération zéros. Finalement, on rétablit les valeurs initiales de X qui normalement être supprimées sont remplacées par des zéros. Pour conserver les dimensions de X , on modifie les opérations de suppression de manière que les valeurs qui devraient demeurer telles quelles après l'application de M . Pour celles que $\text{Ign}(\bar{Q}, R)$ a remplacées par une valeur nulle – sélectionnées, toutes les valeurs "annulées" – c.-à-d. Comme M est destiné à modifier uniquement les valeurs changements voulus, ce qui donne $\tilde{X} = M(\text{Ign}(\bar{Q}, R))$, appropriés M de la section 3 à $\text{Ign}(\bar{Q}, R)$ pour effectuer les on applique le masque ou la combinaison de masques même opération pour les colonnes. À la deuxième étape, substitue des zéros à toutes les autres valeurs; B fait la

ou des *micro-moyennes* de sous-ensembles des enregistrements initiaux.

L'agrégation d'enregistrements peut se faire de plusieurs manières. Une façon classique est de remplacer tous les cumulatifs par les totaux correspondants. Supposons que les enregistrements qui doivent faire l'objet d'une micro-agrégation sont ordonnés et désignons les tailles respectives des groupes d'enregistrements par n_1, n_2, \dots, n_s où $n = n_1 + n_2 + \dots + n_s$. La micro-agrégation peut s'effectuer au moyen d'une matrice A diagonale par blocs de dimension $n \times n$. La diagonale principale de A est formée d'un bloc ordonné de matrices J carrées de dimension $n_v \times n_v$, $v = 1, \dots, s$; les autres éléments de A sont nuls. Dans une micro-agrégation, les valeurs initiales sont remplacées par des micro-moyennes, les valeurs initiales sont remplacées par des micro-moyennes). On peut aussi remplacer un enregistrement, dans chaque groupe, par l'enregistrement ayant fait l'objet d'une micro-agrégation tandis que les autres enregistrements sont supprimés. Cela peut se faire au moyen de matrices J de dimension $1 \times n_v$, auquel cas la dimension de A est $s \times n$. Pour établir des micro-moyennes au lieu de micro-agrégats, on remplace chaque matrice J par son équivalent $1/n_v J$.

3.3 Modification de l'ordre des enregistrements

Le fichier de microdonnées X qui est constitué en vue d'un usage collectif provient habituellement d'un fichier de données plus vaste (par échantillonnage par exemple) ou d'un fichier plus détaillé (à la condition qu'on supprime les données personnelles telles que le nom, l'adresse et le numéro de sécurité sociale). Dans le premier cas, les enregistrements du fichier source sont souvent classés dans un ordre prescrit, par exemple selon la région géographique ou le numéro de sécurité sociale, et X risque fortement de reproduire cet ordre. Pour réduire les risques de divulgation, on doit modifier l'ordre des micro-enregistrements de X . Cette opération peut se faire au moyen d'une matrice A stochastique. Etant donné un réarrangement des lignes (enregistrements) de X (c.-à-d. une permutation P des numéros de ligne $\{1, \dots, n\}$), alors pour $P(i) = h$, posons la i -ième ligne de A égale à la matrice U_{1h} de dimension $1 \times n$. A est désignée par $\text{Reo}(P)$. Une autre formulation est $\text{Reo}(P) = \sum_{i=1}^n U_{1i} p(i)$.

3.4 Arrondissement et perturbation de microdonnées

Les organismes statistiques utilisent l'arrondissement de données à plusieurs fins, notamment pour la protection du secret statistique. Si des variables entières comme l'âge ou le nombre d'années passées sur le marché du travail ou encore le nombre d'enfants étaient reproduites telles quelles, elles pourraient servir, une fois combinées à d'autres informations, à révéler l'identité de répondants (Bethlehem, Keller et Pannekoek 1990). L'arrondissement classique (c.-à-d. arrondissement à un multiple de 5; les valeurs se terminant par 1 ou 2 sont arrondies au multiple de 5 inférieur et les valeurs se terminant par 3 ou 4, arrondies

La matrice I supérieure de $\text{Agg}(f, k)$ est de dimension $(k-1) \times (k-1)$, la matrice I inférieure, de dimension $(p-k) \times (p-k)$, et la matrice U centrale (U_{1j}), de dimension $1 \times (p-1)$. Une autre façon de formuler la matrice B est la suivante:

$$\text{Agg}(f, k) = \text{Supp}(k) + U_{kj}, \quad \text{pour } j < k, \quad \text{et} \\ \text{Agg}(f, k) = \text{Supp}(k) + U_{k, j-1}, \quad \text{pour } j > k.$$

L'agrégation-suppression appliquée à plus de deux attributs peut être représentée comme le produit de matrices B formulées comme ci-dessus. Construisons B_1 comme ci-dessus pour fondre les deux premiers attributs en un sous-total, remplacer le premier attribut par ce sous-total, puis supprimer le deuxième attribut. Procédons de la même manière pour B_2, \dots, B_{c-1} jusqu'à ce que tous les attributs cumulatifs aient été incorporés dans le total, puis supprimés. Alors, $B = B_1 \dots B_{c-1}$.

On peut aussi formuler l'opération d'agrégation-suppression – regroupement des attributs j et k , remplacement de l'attribut j et suppression de l'attribut k – par le produit de matrices $B \text{Add}(j, k) \text{Supp}(k)$. Par ailleurs, il est possible de regrouper les attributs j et k de remplacer l'attribut j sans supprimer l'attribut k en utilisant la matrice B de dimension $p \times p$: $\text{Add}(j, k) = I + U_{kj}$. On peut étendre cette dernière formulation à un plus grand nombre de cumulatifs v en ajoutant des U_{vj} . Pour créer un nouvel attribut totalisateur (attribut $p+1$) à partir des attributs j et k sans devoir remplacer aucun de ces attributs, formons la matrice B de dimension $p \times (p+1)$ [$I \mid U_{j1} + U_{k1}$], où la matrice I est de dimension $p \times p$ et la sous-matrice $p \times 1$. L'introduction d'un autre attribut v dans l'agrégation revient à ajouter des U_{v1} dans la sous-matrice de droite.

Le regroupement de données qualitatives, appelé parfois *regroupement de catégories*, peut être représenté comme l'agrégation d'attributs. Représentons chacune des catégories disjointes d'une variable qualitative, qui sont au nombre de c , par une colonne de X . Chaque colonne contiendra des uns ou des zéros selon que le caractère correspondait ou non. Le regroupement des c catégories en une seule équivalait simplement à une agrégation des c attributs, par laquelle on remplace un attribut par l'agrégat et supprime les autres attributs en utilisant les matrices B de la manière décrite plus haut.

Il est parfois souhaitable d'agréger des valeurs d'attribut relatives à des micro-enregistrements. Par exemple, s'il est possible de grouper des micro-enregistrements selon un critère de "ressemblance" (p. ex., âge ou profession, ou, pour les entreprises d'une industrie en particulier, valeur totale des livraisons ou effectif), alors au lieu de diffuser des micro-enregistrements qui risquent fort de révéler des données confidentielles, on diffuse un fichier de micro-données dont les enregistrements sont des *micro-agrégats*

Un masque de matrice élémentaire de X est un masque de forme AX, XB ou $X + C$. Des itérations de masques de (élémentaires), d'une matrice X sont aussi des masques de cette matrice. Par conséquent, un masque de X a la forme $X = AXB + C$, où X est égal à X ou bien a été déduit de X par l'application d'une suite de masques de matrice élémentaires. Un avantage majeur de cette définition est qu'elle permet d'appliquer sélectivement diverses méthodes de protection du secret statistique à des sous-ensembles d'attributs des enregistrements et des attributs de X arbitraires.

Les matrices A, B, C ne sont pas nécessairement fixes. Par exemple, l'application d'un masque à des attributs numériques comporte souvent l'introduction de bruit aléatoire (Tendick 1991), de sorte que C est une matrice aléatoire. Les matrices A, B, C peuvent dépendre de X . Par exemple, pour "décaler" X au moyen de bruit aléatoire additif proportionnel à la taille, tirons aléatoirement les c_{ij} d'une distribution normale de moyenne nulle et d'écart type égal à un multiple de $|x_{ij}|$ et posons $\tilde{X} = X + C$. Ou bien, si $A = X, M = AX$ est suffisant pour une régression par les moindres carrés ordinaires (Duncan et Pearson 1991).

2.2 Notation

I désigne la matrice identité, Z , la matrice dont tous les éléments sont nuls et J , la matrice composée essentiellement de uns. U_j désigne la matrice dont tous les éléments sont nuls, sauf $u_j = 1$. I est toujours une matrice carrée, alors que ce n'est pas nécessairement le cas pour Z , J et U_j . La matrice U_j conserve les valeurs d'une seule ligne ou d'une seule colonne de la matrice par laquelle elle est multipliée, selon qu'elle sert de pré-multiplicateur ou de post-multiplicateur. La dimension des sous-matrices peut varier d'une formulation à l'autre ou à l'intérieur même d'une formulation et nous en définirons les diverses valeurs pour plus de clarté.

3. REPRÉSENTATION DE MASQUES DE DONNÉES COMME DES MASQUES DE MATRICE ÉLÉMENTAIRES

3.1 Suppression sélective de microdonnées

La première méthode de protection du secret statistique qui nous vient intuitivement à l'esprit est celle qui consiste à soustraire purement et simplement certaines micro-données à la publication. Ces données sont habituellement celles qui impliquent le plus grand risque de divulgation et leur diffusion peut exiger au préalable la suppression d'attributs (colonnes) ou d'enregistrements (lignes) de X . La suppression de l'attribut k peut être représentée comme un masque de transformation d'attribut $X = XB$, où B est la matrice de matrices de dimension $p \times (p - 1)$:

$$\begin{bmatrix} I & Z \\ Z & I \end{bmatrix} = (\gamma) d d n S$$

Supp(k) = $\Sigma_{j < k} U_{jj}^* + \Sigma_{j > k} U_{jj}^{*-1}$.

représentée comme le produit de matrices B formulées comme ci-dessus. Par exemple, $\text{Supp}(k)\text{Supp}(f)$ supprime tout d'abord le k -ième attribut de X , puis supprime le j -ième attribut de la matrice résultante $X\text{Supp}(k)$ de dimension $n \times (d - 1)$. Les dimensions de $\text{Supp}(k)$ et de $\text{Supp}(f)$ sont $d \times (d - 1)$ et $(d - 1) \times (d - 2)$ respectivement. Il est parfois nécessaire de supprimer des enregistréments de la matrice X , soit parce qu'il y a de fortes chances qu'un répondant puisse être identifié, par exemple, ou parce qu'il s'agit d'un enregistrément faux ou inadmissible. La suppression de l'enregistrément h peut être représentée comme un masque de transformation d'enregistrement $\bar{X} = AX$, où A est une matrice de matrices de dimension $(n - 1) \times n$ qui a la même structure que $\text{Supp}(h)$, sauf que la matrice Z centrale est de dimension $(n - 1) \times 1$ et que les dimensions des matrices I supérieure et inférieure sont $(h - 1) \times (h - 1)$ et $(n - h) \times (n - h)$ respectivement. Cette matrice A est désignée par $\text{Del}(h)$. Une autre façon de la formuler est $\text{Del}(h) = \sum_{i < h} U_{hi}^T + \sum_{i > h} U_{hi}^{-1}$.

La suppression de plus d'un enregistrement à la fois est représentée comme le produit de matrices $A \text{ Del}(h)$. Par exemple, pour supprimer les enregistrements h et i de X , on utilise $\text{Del}(i) \text{ Del}(h)$. Si $i > h$, on utilise $\text{Del}(i - 1) \text{ Del}(h)$. Les dimensions de $\text{Del}(i - 1)$ et de $\text{Del}(h)$ sont $(n - 2) \times (n - 1)$ et $(n - 1) \times n$ respectivement. La matrice A qui *supprime systématiquement* un enregistrement à tous les h enregistrements (pour $n = rh$, r étant un entier) est une matrice de matrices comprenant r matrices $\text{Del}(h)$ de dimension $(h - 1) \times n$ disposées à la verticale. Cette définition s'étend à la suppression non systématique.

Le complètement de la suppression d'enregistrements est l'échantillonnage *d'enregistrements*. La matrice A qui échantillonne systématiquement un enregistrement de X à tous les h enregistrements (pour $n = nh$) est une matrice $r \times n$ dont la q -ième ligne est la matrice $1 \times n U_{1,qh}$. D'une manière plus générale, pour tirer un échantillon de tailles s formé des enregistrements de X identifiés par l'ensemble $S = \{s_p : p = 1, \dots, s\}$, on utilise la matrice $A \text{ Sam}(X, S)$ de dimension $s \times n$, dont chaque ligne est une matrice U_{1,s_p} de dimension $1 \times n$.

3.2 Regroupement de microdonnées

Plus les données sont agrégées, moins il y a de chances qu'un répondant puisse être identifié ou que des données confidentielles soient divulguées. L'agrégation d'attributs et d'autres masques de microdonnées reposent sur ce principe. Le masque d'agrégation qui remplace le premier de deux attributs (l'attribut j) par la somme de ces attributs et supprime le second (l'attribut k) de la matrice X , pour $j < k$, peut être représenté comme une transformation d'attribut $\tilde{X} = XB$, où B est la matrice de matrices de dimension $p \times (p - 1)$:

Méthodes de masquage de matrice pour la protection du caractère confidentiel de microdonnées

LAWRENCE H. COX¹

RÉSUMÉ

De nombreuses méthodes de protection du caractère confidentiel des microdonnées sont décrites dans les ouvrages de statistique. Cependant, l'usage qu'en font les organismes statistiques et la compréhension qu'on a de leurs propriétés et de leurs effets sont limités. Afin de favoriser la recherche sur ces méthodes ainsi que leur usage et pour faciliter leur évaluation et l'assurance de la qualité, il est souhaitable de formuler ces méthodes selon une seule approche. Dans cet article, nous présentons une approche appelée *masquage de matrice* – qui repose sur le calcul matriciel ordinaire – et nous formulons des masques de matrice pour les principales méthodes de protection du caractère confidentiel de microdonnées actuellement en usage, ce qui permettra aux organismes statistiques et aux autres spécialistes du domaine d'avoir une meilleure compréhension de ces méthodes et de les mettre en application.

MOTS CLÉS: Protection du secret statistique; traitement des données d'enquête; méthodes mathématiques.

1. INTRODUCTION

À l'ère de l'information, les données sont devenues un élément indispensable au fonctionnement d'institutions qui jouent un rôle primordial dans notre société. Les utilisateurs de données statistiques comptent particulièrement sur les organismes gouvernementaux de statistique pour recueillir des données fiables et les diffuser dans les meilleurs délais sous des formes les plus variées. Avant les années 1950, les données étaient diffusées uniquement sous forme de tableaux imprimés. Dans les années 1960, le gouvernement des États-Unis a commencé à publier des données individuelles (*microdonnées statistiques*).

À l'heure actuelle, les chercheurs et les analystes de la politique qui ne font pas partie d'organismes statistiques ont beaucoup de difficulté à obtenir des microdonnées pour leurs travaux parce qu'on leur refuse l'accès aux données voulues pour des raisons de confidentialité. Depuis une trentaine d'années, les organismes statistiques se débattent au milieu de problèmes d'ordre général ou technique concernant la publication de microdonnées, et plusieurs de ces problèmes sont encore irrésolus (Federal Committee on Statistical Methodology 1994). Le but de cet article est d'exposer une série de transformations matricielles appliquées à des microdonnées qui devraient aider les organismes statistiques à résoudre ces difficultés.

Duncan (1990) et Duncan et Pearson (1991) ont défini plusieurs méthodes de protection du secret statistique pour microdonnées (*masques de microdonnées*) qui reposent sur l'addition et la multiplication de matrices, et ils ont appelé ces méthodes des "masques de matrice". Cox (1991) a généralisé la notion de masque de matrice et a étendu la définition à d'autres masques de microdonnées. Le fait de définir les masques de microdonnées comme des masques de matrice présente des avantages sur le plan théorique et le plan statistique. Le masquage de matrice permet l'utilisation d'un

langage simple pour représenter, comparer et évaluer les méthodes de masquage de microdonnées. Cette approche permet d'exprimer des méthodes variées et compris les statistiques et les utilisateurs de données, et offre une structure normalisée pour le développement de logiciels de masquage interchangeables et l'optimisation de leur performance. Dans cet article, la notion de masque de matrice est traitée de façon mathématique. Nous formulons des masques de matrice pour les principales méthodes de masquage de microdonnées en usage actuellement et nous étendons en même temps la portée des masques présentes dans Duncan et Pearson (1991) et Cox (1991), de sorte que ces méthodes pourront être appliquées facilement sous forme logicielle et que les organismes statistiques pourront étudier de plus près les masques de microdonnées et en faire l'utilisation. Ainsi, on devrait pouvoir mieux comprendre les propriétés des masques de microdonnées et, surtout, l'incidence de ces masques sur l'utilisation des données.

2. MASQUES DE MATRICE

2.1 Définitions

Un fichier de microdonnées contenant p valeurs d'attribut pour chaque enregistré comme une matrice X de dimension $n \times p$ dont les éléments sont désignés par x_{ij} . À moins d'indication contraire, X ne renferme aucune valeur manquante. Un *masque de matrice* (A, B, C) est une transformation de X de forme $X' = AXB + C$, avec $A, B \neq 0$, qui implique l'addition et la multiplication ordinaires de matrices. Comme A opère sur les lignes de X , elle est appelée *masque de transformation d'enregistrement*. B est un *masque de transformation d'attribut* et C , un *masque de décalage* ("displacing mask") (Duncan et Pearson 1991).

¹ Lawrence H. Cox, Senior Statistician, United States Environmental Protection Agency, AREAL (MD-75), Research Triangle Park, NC 27711, U.S.A.

BIBLIOGRAPHIE

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhya*, C, 37, 117-132.

KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, 107-112.

RAO, J.N.K., et BELLHOUSE, D.R. (1989). The history and development of the theoretical foundations of survey based estimation and statistical analysis. *American Statistical Association Proceedings Sesquicentennial Invited Paper Sessions*, 406-428.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SATTERTHWAITTE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

SHAH, B.V., HOLT, M.M., et FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin de l'Institut International de Statistique*, 47, 43-57.

SKINNER, C.J. (1989). Domain means, regression, and multivariate analysis. Dans *Analysis of Complex Surveys* (Eds. C.J. Skinner, D. Holt et T.M.F. Smith). New York: John Wiley, 59-88.

WU, C.J.F., HOLT, D., et HOLMES, D.J. (1988). The effect of two stage sampling on the *F* statistic. *Journal of the American Statistical Association*, 83, 150-159.

REMERCIEMENTS

L'auteur tient à remercier le personnel du Beltsville Human Nutrition Research Center pour son soutien à ce travail de recherche et un rédacteur associé et ses arbitres pour leurs commentaires utiles.

On peut illustrer avec force la question de la puissance en reprenant l'exemple simple de la section 6. L'estimation basée sur un modèle et l'estimation basée sur un plan sont identiques. Si l'on suppose que tous les ϵ_i sont identiquement distribués, l'estimateur de la variance fondé sur un modèle, v_M , qui dépend de l'hypothèse d'homoscédastité, est non biaisé et a 98 degrés de liberté. La version corrigée de l'estimateur de la variance fondé sur un plan est aussi virtuellement non biaisée, mais elle n'a que 9 degrés de liberté.

En pratique, il sera souvent sage de sacrifier la puissance à la robustesse. L'équation (6) offre alors un moyen intéressant d'évaluer la perte probable de puissance quand on utilise le test *t* modifié (équation (7)) et que les hypothèses du modèle sont justes. En outre, cette même équation se prête à des analyses de sensibilité par lesquelles on peut mesurer les effets de diverses hypothèses concernant les v_{h_j} .

où $A = 2XCes\epsilon_j - XCes\epsilon_j'CX'$. On peut maintenant

montrer que

$$s_2^{*2} = n^{-2} \sum_{h=1}^H (n_h/[n_h - 1]) \sum_{n_h} (e_{hj} - e_h)^2$$

$$+ O_p(n^{-5/2}).$$

Considérons une variable aléatoire qui suit une distribution χ^2 avec F degrés de liberté. Sa variance relative est $2/F$. Cela évoque une estimation de type Satterthwaite du nombre effectif de degrés de liberté de s_2^2 (voir Satterthwaite 1946), c'est-à-dire

$$F = \frac{\sum_{h=1}^L \left\{ \sum_{j=1}^{n_h} v_{hj}^4 + \sum_{j' \neq j} v_{hj}^2 v_{hj'}^2 / (n_h - 1)^2 \right\}}{(nv)^4}, \quad (6)$$

ce qui est le quotient de 2 environ par la variance relative de s_2^{*2} (puisque $s_2^{*2} \approx n^{-2} \sum_h \{ \sum_j e_{hj}^2 + \sum_{j' \neq j} e_{hj} e_{hj'} / (n_h - 1) \}$).

La voie à suivre dans les circonstances serait de tester l'hypothèse que $qg = \Theta_0$ en supposant selon l'hypothèse

nuile que

$$t^* = (qb^w - \Theta_0)/s^*, \quad (7)$$

suit une distribution t de Student avec F degrés de liberté, F étant calculé au moyen de l'équation (6) et des hypothèses étant faites à propos des v_{hj} . Appelons ce test le

test t modifié.

6. UN EXEMPLE SIMPLE

Considérons un échantillon aléatoire simple de n unités, dont n_1 forment le sous-ensemble désigné par A et n_2 l'autre sous-ensemble, désigné par \bar{A} . Soit y_i la valeur observée pour l'unité i . Supposons que le modèle linéaire suivant est valide:

$$y_i = d_i \beta_1 + (1 - d_i) \beta_2 + \epsilon_i, \quad (8)$$

où d_i égale 1, si l'unité i appartient au sous-ensemble A , et 0, si elle appartient à l'autre sous-ensemble, et où les ϵ_i sont des variables aléatoires indépendantes distribuées normalement.

En supposant l'homoscédasticité des erreurs, nous avons comme estimateur par régression de β_1 (en version modèle et en version plan) la moyenne de domaine non pondérée $\bar{y}_A = \sum_{i \in A} y_i / n_1$. L'estimateur par linéarisation de la variance de cet estimateur est simplement $v_L = (n/[n - 1]) \sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1^2$. (Notons que lorsqu'une moyenne de domaine est considérée comme un paramètre analytique, sa variance ne nécessite pas de correction pour population finie; voir Fuller 1975.)

$$t^* = \frac{\sum_{i \in A} y_i / n_1 - \sum_{i \in \bar{A}} y_i / n_2}{(1 - s^{-2} R_{1/2}^2)^{1/2}},$$

où

$$s^2 = [n/(n - 1)] [\sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1^2$$

et

$$R = [n/(n - 1)] [(\sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1^3) (1 - n_1/n) + (\sum_{i \in \bar{A}} (y_i - \bar{y}_A)^2 / n_2^3) (1 - n_2/n)].$$

Il serait normal de tester l'hypothèse que les moyennes 97 ou 99 degrés de liberté.

plus que dans le cas précédent mais nettement mieux que obtient le chiffre de 9.99, ce qui est presque un degré de de degrés de liberté, F , au moyen de l'équation (6), on sous-ensembles A et \bar{A} et en calculant le nombre effectif que les erreurs sont identiquement distribuées dans les l'écart entre s_2^{*2} et v_L n'est que de 0.170. En supposant presque identique à v_L (comme $R = [v_L/n_1] [1 - n_1/n]$), linéarisation, v_L , on obtient un estimateur de la variance Si on combine l'équation (5) à l'estimateur de variance par avec 9 degrés de liberté seulement.

t calculée au moyen de v_L suit une distribution t de Student homoscédastiques dans le sous-ensemble A , la statistique effectuée). Or, suivant des conditions idéales (erreurs catives, cette dernière soustraction n'étant pas toujours d'échantillonnage moins 1 strate moins 2 variables explicatives, cette dernière soustraction n'étant pas toujours de la variance biaisée, mais elle supposerait aussi que la statistique a 97 ou 99 degrés de liberté (c.-à-d. 100 unités sur un plan, non seulement elle utiliserait un estimateur une statistique t à l'aide de la méthode classique fondée Revenons à notre exemple. Si une personne calculait parfaitement non biaisé.

$(n_1[n_1 - 1]) = ([n - 1]/n)(n_1/[n_1 - 1])v_L$ est par conséquent, il suffit de remarquer que $v_L = \sum_{i \in A} (y_i - \bar{y}_A)^2 / n_1 = 10$, le biais relatif de v_L est d'environ 10%. Pour le pour un nombre n fini. Par exemple, lorsque $n = 100$ et ment, on peut néanmoins observer un biais appréciable relevée par Skinner (1989) et Kott (1991). Malheureusement, sont hétéroscédastiques. Cette caractéristique a été L'avantage de v_L par rapport à v_M est qu'il est asymptotiquement non biaisé selon le modèle même lorsque les L'estimateur par linéarisation v_L diffère de l'estimateur de la variance fondé sur un modèle, soit $v_M = [\sum_{i \in A} (y_i - \bar{y}_A)^2 + \sum_{i \in \bar{A}} (y_i - \bar{y}_A)^2] / [n_1(n - 2)]$.

en (1) et de l'hypothèse selon laquelle Σ_S est une matrice diagonale par blocs. Nous procéderons en analysant $s^2 = q\text{Var}(b^w)q'$. Premièrement, nous corrigerons s^2 pour en réduire le biais; ensuite, nous déterminerons une meilleure méthode pour calculer le nombre effectif de degrés de liberté de l'estimateur ainsi corrigé.

4. LE BIAIS DE MODELE DE s^2

L'analyse que nous allons faire est asymptotique. Bon nombre des résultats dépendent de l'hypothèse que $n - \text{le nombre d'unités primaires d'échantillonnage dans l'échantillon} - \text{est grand}$. (Du point de vue de la forme, nous devrions supposer qu'il existe une suite infinie de statistiques qui prennent une valeur lorsque n devient arbitrairement élevée.) Si n est grand, il doit en être de même pour M et m , c'est-à-dire pour l'effectif de la population et de l'échantillon, respectivement. Nous supposons que $\max\{m_h\}$ est borné par une valeur finie, disons m_0 . Donc, m est borné par $m_0 n$ et le nombre d'éléments non nuls dans la matrice diagonale par blocs Σ_S est borné par $m_0 n$.

Le nombre de colonnes dans X_S , K , est supposé fixe, mais, pour le nombre de strates, L , nous avons une certaine latitude. Soit que L demeure fixe lorsque n tend vers une valeur arbitrairement élevée, les rapports n_h/n tendant vers une limite positive fixe, soit que L/n tende vers une limite positive fixe, $\max\{n_h\}$ étant borné.

L'important, c'est que nous définissions des conditions *suffisantes* pour que l'analyse qui va suivre soit valide. Disons que la variable aléatoire ϕ (formellement, la suite aléatoire infinie $\{\phi_n\}$) est d'ordre de probabilité $n^{-\delta}$, c.-à-d. $\phi = O_p(n^{-\delta})$ lorsque $|E(\phi^2)| < B/n^{2\delta}$ pour une valeur B finie. De même, nous dirons que la matrice aléatoire Φ est égale à $O_p(n^{-\delta})$ lorsque chaque élément ϕ_{ij} dans Φ satisfait l'inéquation $|E(\phi_{ij}^2)| < B/n^{2\delta}$. Lorsque ϕ n'est pas aléatoire, il n'est pas nécessaire d'affecter O de l'indice P . Même chose pour O . Les hypothèses suivantes sont raisonnables, compte tenu de la structure qui a été définie:

$$(1) C = (X'WX)^{-1}X'W \text{ existe et est } O(1/n);$$

$$(2) E(\Sigma_{hj}) = \Sigma_{hj} + O(1/n), \text{ où } \Sigma_{hj} = r_{hj}r'_{hj}.$$

L'hypothèse 1 atteste que $\text{Var}(b^w) = C\Sigma_S C' = O(1/n)$ puisque les lignes de C comptent m éléments et que Σ_S contient tout au plus $m_0 n$ éléments non nuls.

On peut récrire la variance de qb^w sous la forme $v^2 = \Sigma v_{hj}/n^2$, où $v_{hj} = n^2 g_{hj} \Sigma_S g_{hj}' = qCD_{hj}$ et D_{hj} est une matrice diagonale composée de uns (pour les éléments échantillonnés de l'u.p.é. h_j) et de zéros (pour les autres). De même, on peut récrire $s^2 = qeqmq'$ sous la forme

$$s^2 = \sum_L \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{n_h} (g_{hj} - g_h) r_S r_S' (g_{hj} - g_h)' \quad (4)$$

$$= \sum (n_h/[n_h - 1]) \sum [g_{hj} \Sigma_S g_{hj}'$$

$$- 2g_h \Sigma_S g_h' + g_h \Sigma_S g_h'],$$

où $g_h = \Sigma g_{hj}/n_h$, la sommation étant étendue à tous les j dans h , et $\Sigma_S = \Sigma \Sigma D_{hj} r_S r_S' D_{hj}'$.

Les vecteurs g_{hj} et g_h sont tous deux $O(1/n)$ parce que $C = O(1/n)$ et que D_{hj} contient un nombre limité d'éléments non nuls. Donc, $E(g_{hj} \Sigma_S g_{hj}') = g_{hj} \Sigma_S g_{hj}' + O(n^{-3})$, $E(g_h \Sigma_S g_h') = g_h \Sigma_S g_h' + O(n^{-3})$, et $E(g_h \Sigma_S g_{hj}') = g_h \Sigma_S g_{hj}' + O(n^{-3})$. Par conséquent, $E(s^2 - v^2) = O(n^{-2})$.

Comme $r_S = (I_m - XC)\epsilon_S$ et $E(\epsilon_S \epsilon_S') = \Sigma_S$, $E(r_S r_S') = \Sigma_S - XC\Sigma_S - \Sigma_S C'X' + XC\Sigma_S C'X'$. Nous pouvons voir d'après l'équation (4) que $E(s^2) = v^2 - R$, où $R = \Sigma (n_h/[n_h - 1]) \Sigma (g_{hj} - g_h) Z(g_{hj} - g_h)'$ et $Z = 2XC\Sigma_S - XC\Sigma_S C'X'$. Or $Z = O(1/n)$, parce que $C = O(1/n)$, X a un nombre fixe de colonnes et le nombre de termes non nuls que contient chaque colonne de Σ_S est limité, ce qui implique que $R = O(n^{-2})$. Par conséquent, $-R/v^2$, le biais relatif de s^2 , est $O(1/n)$.

Un autre estimateur de v^2 , qui a un biais relatif moindre, est

$$s^2_* = s^2/(1 - s^{-2}R), \quad (5)$$

où

$$R = \left\{ \sum_L \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{n_h} (g_{hj} - g_h) Z(g_{hj} - g_h)' \right\},$$

et

$$Z = 2XC\Sigma_S - XC\Sigma_S C'X'.$$

Dans l'équation (5), R/s^2 sert à estimer R/v^2 . Si nous proposons ici l'estimateur de la variance s^2_* au lieu de l'estimateur plus courant $s^2 + R$, c'est uniquement parce que R , en tant qu'estimateur de R , a un biais relatif non négligeable.

5. LA VARIANCE RELATIVE DE L'ESTIMATEUR DE LA VARIANCE

Posons $e_{hj} = n g_{hj} \epsilon_S$, de sorte que $\text{Var}(e_{hj}) = v_{hj}^2$, et rappelons que $v^2 = \Sigma v_{hj}^2/n^2$. Si $\hat{e}_{hj} = n g_{hj} r_S$, la variable aléatoire s^2 peut être reformulée comme suit:

$$s^2 = n^{-2} \sum_L \sum_{h=1}^L (n_h/[n_h - 1]) \sum_{n_h} (\hat{e}_{hj} - \hat{e}_h)^2$$

$$= n^{-2} \Sigma (n_h/[n_h - 1]) \{ \Sigma (e_{hj} - e_h)^2$$

$$- (g_{hj} - g_h) A(g_{hj} - g_h)' \},$$

nous supposons que la population est divisée en L strates. Dans chaque strate h , nous tirons aléatoirement n_h grappes d'éléments *distinctes* et nous les désignons par $u_{h1}, u_{h2}, \dots, u_{hn_h}$. Ensuite, nous tirons dans chaque grappe h_j un échantillon aléatoire de m_{hj} éléments. Les grappes sont aussi appelées unités primaires d'échantillonnage. L'échantillon compte en tout $n = \sum n_h$ unités primaires d'échantillonnage.

Chaque élément de l'échantillon est désigné par les lettres h_ji , h étant la strate dans laquelle se trouve l'élément, h_j l'unité primaire d'échantillonnage dans h et i l'élément proprement dit dans h_j . Soit p_{hji} la probabilité d'échantillonnage de l'élément h_ji et $w_{hji} = m / (Mp_{hji})$, le poids d'échantillonnage de l'élément. Notons que les poids d'échantillonnage ont été normalisés, de sorte que si p_{hji} égale la fraction de sondage, m/M , w_{hji} sera égal à un. Le modèle linéaire défini en (1) vaut aussi pour les éléments de l'échantillon S :

$$y_S = X_S \beta + \epsilon_S,$$

où y_S , par exemple, est le vecteur $m \times 1$ des valeurs de l'échantillon pour la variable dépendante. Soit $\epsilon_{hj} = (\epsilon_{hj1}, \epsilon_{hj2}, \dots, \epsilon_{hjm_{hj}})$ le vecteur d'erreurs pour les éléments de l'unité primaire d'échantillonnage h_j . Or il est possible de définir le vecteur ϵ_S de sorte que les ϵ_{hj} soient disposés les uns au-dessus des autres. Désignons $\text{Var}(\epsilon_{hj}) = E(\epsilon_{hj}\epsilon_{hj}')$ par la matrice $m_{hj} \times m_{hj}$ Σ_{hj} , qui n'est pas nécessairement diagonale. Nous supposons que les ϵ_{hj} ne sont pas corrélés d'une u.p. à l'autre, de sorte que Σ_S est une matrice diagonale par blocs.

L'estimateur de β fondé sur un plan est l'estimateur par les moindres carrés pondérés:

$$b_W = (X_S' W X_S)^{-1} X_S' W y_S,$$

où W est la matrice diagonale $m \times m$ des poids d'échantillonnage. Le g ème élément de la diagonale de W est le poids d'échantillonnage rattaché au g ème élément de l'échantillon. De toute évidence, b_W est un estimateur non biaisé de β selon le modèle défini en (1).

On peut simplifier l'expression de b_W en posant $C = C_W$ comme la matrice $k \times m$ $(X_S' W X_S)^{-1} X_S' W$, de sorte que $b_W = C y_S$. Soit D_{hj} , une matrice diagonale $m \times m$ com-posée de uns (pour les éléments de h_j échantillonnés) et de zéros (pour les autres). En outre, posons $C_{hj} = C D_{hj}$. Finalement, soit $r_S = y_S - X_S b_W$, le vecteur des résiduels. L'estimateur par série de Taylor, ou par linéarisation, de l'erreur quadratique moyenne de b_W (Shah et coll. 1977) est

$$\text{eqm} = \sum_{h=1}^L (n_h / [n_h - 1]) \sum_{h_j} A_{hj} r_S' r_S' A_{hj}' \quad (2)$$

où $A_{hj} = C_{hj} - n_h^{-1} \sum C_{hg}$, la sommation étant étendue à toutes les unités primaires d'échantillonnage de la strate h . Les termes "série de Taylor" et "linéarisation" signifient que l'eqm est calculée selon la théorie des sondages fondée sur un plan. Kott (1991) montre que eqm est un estimateur quasi non biaisé de la *variance* de modèle de b_W suivant des conditions raisonnables.

Il convient de souligner que dans leur calcul de eqm, Shah et coll. ont supposé que les unités primaires d'échantillonnage étaient tirées avec remise. Dans la présente analyse, comme dans Kott (1991), nous supposons que les u.p.é. sont distinctes, ce qui laisse entendre qu'elles ont été prélevées *sans* remise. Cette différence s'explique par le fait que les deux théories – celle fondée sur un plan et celle fondée sur un modèle – supposent des conditions presque contraires pour l'indépendance des u.p.é. échantillonnées à l'intérieur d'une strate. Toutefois, la différence disparaît si nous supposons que les u.p.é. ont été tirées sans remise mais que le but de la théorie de la régression fondée sur un plan est d'estimer non le paramètre d'une population finie mais la valeur limite de ce paramètre lorsque l'effectif de la population (et le nombre d'unités primaires d'échantillonnage par strate) atteint un niveau arbitrairement élevé. Voir Fuller (1975).

Si le modèle défini en (1) est valide et que $L > 1$, il existe un autre estimateur de l'erreur quadratique moyenne de b_W qui est aussi quasi non biaisé. Il a la même forme que l'estimateur de l'équation (2), sauf que l'on fait comme si les n u.p.é. échantillonnées venaient toutes de la même strate ($L = 1$). Comme ce second estimateur peut être exprimé au moyen de l'équation (2), il n'est pas nécessaire de le considérer séparément dans l'analyse qui suit.

3. UNE STATISTIQUE CLASSIQUE BASÉE SUR UN PLAN

L'estimateur b_W est un K -vecteur. Dans cette section, nous nous intéressons à la statistique t qui sert à tester l'hypothèse unidimensionnelle que $q\beta = \theta_0$ pour un vecteur ligne à K éléments $q = (q_1, q_2, \dots, q_K)$. Le cas d'application le plus courant pour ce type d'hypothèse est celui où l'on cherche à établir si un élément particulier de $\beta = (\beta_1, \dots, \beta_K)$, par exemple β_K , a une valeur nulle. Toutes les valeurs q_i seraient ici nulles, sauf q_K , qui serait égale à 1; θ_0 aurait aussi une valeur nulle.

Si le modèle défini en (1) est valide et que l'hypothèse nulle ($q\beta = \theta_0$) est vraie, alors

$$\Theta = (qb_W - \theta_0) / \{q \text{Var}(b_W) q'\}^{1/2}$$

suivrait une distribution normale de moyenne nulle et de variance 1. Si l'on connaissait $\text{Var}(b_W)$, on pourrait tester l'hypothèse nulle en comparant la statistique Θ aux valeurs d'une table de la distribution normale centrée et réduite. Malheureusement, $\text{Var}(b_W)$ doit être estimée au moyen de l'échantillon. La méthode classique (fondée sur un plan) consiste à comparer la statistique

$$t = (qb_W - \theta_0) / (q \text{eqm} q')^{1/2}, \quad (3)$$

à une distribution t de Student avec $n - L$ ou $(n - L - K)$ degrés de liberté (voir Shah et coll. 1977). Le principal objectif de cet article est d'examiner, puis de modifier, la méthode décrite ci-dessus, qui manque peut-être de généralité, en nous servant du modèle défini

Test d'hypothèse portant sur des coefficients de régression linéaire et basé sur des données d'enquête

PHILLIP S. KOTT¹

RÉSUMÉ

L'objet de cet article est de tester une hypothèse concernant des coefficients de régression linéaire en se fondant sur des données d'enquête. L'article montre que si l'on utilise l'estimateur de variance par linéarisation fondé sur un plan pour un coefficient de régression, il faut modifier cet estimateur pour en réduire le faible biais de modèle et faire une estimation de type Satterthwaite du nombre effectif de ses degrés de liberté. Un aspect particulier très important de cette analyse est son application aux moyennes de domaine.

MOTS CLÉS: Fondé sur un plan; moyenne de domaine; nombre effectif de degrés de liberté; fondé sur un modèle; ordre de probabilité.

1. INTRODUCTION

La théorie statistique est en majeure partie de nature analytique. On a au départ un ensemble de données et un modèle stochastique assez général qui est censé avoir produit ces données. On recourt alors à la théorie statistique pour estimer les paramètres du modèle et déterminer la précision des estimations. En fin de compte, le modèle initial peut se réduire au résultat d'une série de tests statistiques qui souvent consiste à déterminer si l'on peut raisonnablement inférer que les valeurs de certains paramètres sont nulles.

La théorie des sondages est, elle, avant tout descriptive plutôt qu'analytique. L'objet d'étude est une population finie. En principe, l'information relative à cette population peut être résumée par une ou plusieurs statistiques descriptives (par exemple la moyenne et la médiane de la population). À cause des délais qui lui sont impartis ou de contraintes budgétaires, le statisticien d'enquête doit se contenter d'un échantillon de la population pour estimer ces statistiques. La plupart du temps, deux tâches attendent le statisticien dans les circonstances: il doit d'abord choisir une méthode d'échantillonnage; ensuite, il doit estimer les statistiques de la population à partir de l'échantillon formé. Bien qu'il soit possible d'élaborer une théorie statistique fondée sur un modèle pour appuyer des opérations de ce genre (voir, par exemple, Royall 1970), la plupart des statisticiens d'enquête préfèrent une approche sans modèle dite théorie des sondages fondée sur un plan. Selon cette théorie, ce ne sont pas les valeurs de l'échantillon qui sont stochastiques (comme c'est le cas pour la théorie fondée sur un modèle) mais le processus d'échantillonnage. L'ouvrage de Rao et Bellhouse (1989) résume bien les deux approches (celle fondée sur un plan et celle fondée sur un modèle) ainsi que les efforts qui ont été faits pour les combiner.

L'objet de cet article est de tester une hypothèse concernant des paramètres de régression linéaire. Nous supposons que le modèle est correct et que les erreurs du modèle sont

distribuées normalement avec une structure de covariance pouvant être complexe. Contrairement à Wu et coll. (1988), nous ne modéliserons pas explicitement la structure d'erreur (sauf peut-être à un moment ultérieur). Nous allons plutôt tourner notre attention vers une statistique t calculée au moyen de l'estimateur de variance par linéarisation. Skinner (1989) et Kott (1991) ont démontré que cet estimateur a des propriétés de robustesse intéressantes du point de vue de l'approche fondée sur un modèle.

Nous proposerons aussi des méthodes pour réduire le biais de modèle de l'estimateur de variance par linéarisation et calculer le nombre effectif de degrés de liberté de cet estimateur. Un aspect très important de l'analyse sera l'étude de la variance estimée des moyennes de domaine et de la différence de ces moyennes.

Comme l'analyse présentée dans cet article s'inspire essentiellement de l'approche fondée sur un modèle, les termes "biais" et "variance" désigneront respectivement le biais de modèle et la variance de modèle, sauf indication contraire.

2. LE MODÈLE

Supposons que nous ayons une population de M éléments qui peut être décrite par le modèle linéaire:

$$y_M = X_M \beta + \epsilon_M, \quad (1)$$

où y_M est un vecteur $M \times 1$ de valeurs de la population pour la variable dépendante étudiée; X_M est une matrice $M \times K$ de valeurs de la population pour les K variables indépendantes désignées; β est un vecteur $K \times 1$ de coefficients de régression; ϵ_M est un vecteur aléatoire qui suit une distribution normale de moyenne 0_M et de variance Σ_M .

Un échantillon aléatoire, S , formé de m éléments distincts est prélevé dans la population. Pour favoriser un certain degré de généralité dans le plan d'échantillonnage,

¹ Phillip S. Kott, National Agricultural Statistics Service, 3201 Old Lee Highway, Fairfax (VA) 22030, U.S.A.

- MULRY, M.H., et SPENCER, B.D. (1991). Total error in PES estimates of population: the dress rehearsal census of 1988 (avec discussion). *Journal of American Statistical Association*, 86, 839-854.
- MULRY, M.H., et SPENCER, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of American Statistical Association*, 88, 1080-1091.
- RAO, C.R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, 18, 139-148.
- SANATHANAN, L. (1972). Estimating the size of a multi-nomial population. *Annals of Mathematical Statistics*, 43, 142-152.
- SCHENKER, N. (1988). Traitement des données manquantes dans l'estimation de la couverture: le test des opérations de redressement de 1986. *Techniques d'enquête*, 14, 93-104.
- WOLTER, K. (1986). Some coverage error models for census data. *Journal of American Statistical Association*, 81, 338-346.

inférence les effectifs par case $\{y_{ij}^*\}$, $i, j = 1, 2$ et $ij \neq 22$, en se servant des données non pondérées de l'échantillon P et des enregistrements du recensement corrigés par un facteur de k . Alors, $x_{ij} = ky_{ij}$, $ij \neq 22$, et $x_{11} + x_{21} = ky_{11} +$
 $x_{21} = ky_{11}$. Désignons par \hat{p}_1, \hat{p}_2 et N_w les estimations de système dual tirées de $\{x_{ij}^*\}$ et par \hat{q}_1, \hat{q}_2 et N_u les esti-
 mations tirées de $\{y_{ij}^*\}$. Nous avons (Bishop et coll. 1975, chap. 6) $\hat{p}_1 = x_{11}/x_{+1} = y_{11}/y_{+1} = \hat{q}_1, \hat{p}_2 = x_{11}/x_{1+} =$
 $y_{11}/y_{1+} = \hat{q}_2, N_w = x_{1+} + x_{21}/x_{11} = ky_{1+} + y_{21}/y_{11} = kN_u$.
 Donc, si l'on opte pour les données non pondérées de l'échantillon P et qu'on utilise la formule $N^* = kN_u$ pour
 estimer le total de la population, on obtiendra les mêmes
 estimations ponctuelles avec \hat{q}_1, \hat{q}_2 et N^* qu'avec \hat{p}_1, \hat{p}_2 et
 N_w , celles-ci basées sur les données pondérées de l'échan-
 tillon P. De la distribution normale asymptotique des esti-
 mations (Ding 1993b) nous déduisons: $\text{Var}(N_w) =$
 $k\text{Var}(N_u), \text{Var}(\hat{q}_1) = k\text{Var}(\hat{p}_1), \text{Var}(\hat{q}_2) = k\text{Var}(\hat{p}_2)$.
 Alors, $\text{Var}(N^*) = k\text{Var}(N_u)$, et \hat{q}_1, \hat{q}_2 et N^* ont une
 variance plus élevée que \hat{p}_1, \hat{p}_2 et N_w respectivement.
 Pour calculer des estimations avec des données non pon-
 dérées de l'échantillon P, il faut connaître k et $\{y_{ij}^*\}$.
 Nous tenons à souligner que l'hypothèse du poids d'échan-
 tillonnage unique pour toutes les unités d'une strate n'est
 posée ici que pour simplifier l'analyse. Dans la réalité, la
 situation peut être plus complexe. Par exemple, des Noirs
 peuvent être échantillonnés avec une faible probabilité
 dans une strate à majorité blanche, puis être classés avec
 d'autres Noirs qui, eux, sont échantillonnés avec une pro-
 babilité beaucoup plus grande.

BIBLIOGRAPHIE

- inférence les effectifs par case $\{y_{ij}^u\}$, $i, j = 1, 2$ et $u \neq 22$, en se servant des données non pondérées de l'échantillon P et des enregistrements du recensement corrigés par un facteur de k . Alors, $x_{ij}^u = ky_{ij}^u$, $u \neq 22$, et $x_{1+} = ky_{1+}$, $x_{+1} = ky_{+1}$. Désignons par p_1, p_2 et N_w les estimations de système dual tirées de $\{x_{ij}^u\}$ et par \hat{q}_1, \hat{q}_2 et \hat{N}_w les estimations tirées de $\{y_{ij}^u\}$. Nous avons (Bishop et coll., 1975, chap. 6) $\hat{p}_1 = x_{1+}/x_{+1} = y_{1+}/y_{+1} = \hat{q}_1$, $\hat{p}_2 = x_{11}/x_{1+} = y_{11}/y_{1+} = \hat{q}_2$, $N_w^* = x_{1+}x_{+1}/x_{11} = ky_{1+}y_{+1}/y_{11} = kN_w$. Donc, si l'on opte pour les données non pondérées de l'échantillon P et qu'on utilise la formule $N_w^* = kN_w$ pour estimer le total de la population, on obtiendra les mêmes estimations ponctuelles avec \hat{q}_1, \hat{q}_2 et N_w^* qu'avec \hat{p}_1, \hat{p}_2 et N_w , celles-ci basées sur les données pondérées de l'échantillon P. De la distribution normale asymptotique des estimations (Ding 1993b) nous déduisons: $\text{Var}(N_w^*) = k\text{Var}(N_w)$, $\text{Var}(\hat{q}_1) = k\text{Var}(\hat{p}_1)$, $\text{Var}(\hat{q}_2) = k\text{Var}(\hat{p}_2)$. Alors, $\text{Var}(N_w^*) = k\text{Var}(N_w)$, et \hat{q}_1, \hat{q}_2 et N_w^* ont une variance plus élevée que \hat{p}_1, \hat{p}_2 et N_w respectivement. Pour calculer des estimations avec des données non pondérées de l'échantillon P, il faut connaître k et $\{y_{ij}^u\}$. Nous tenons à souligner que l'hypothèse du poids d'échantillonnage unique pour toutes les unités d'une strate n'est posée ici que pour simplifier l'analyse. Dans la réalité, la situation peut être plus complexe. Par exemple, des Noirs peuvent être échantillonnés avec une faible probabilité dans une strate à majorité blanche, puis être classés avec d'autres Noirs qui, eux, sont échantillonnés avec une probabilité beaucoup plus grande.
- ## BIBLIOGRAPHIE
- ALHO, J.M., MULRY, M.H., WURDEMAN, K., et KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- BELIN, T.R., DIFFENDAL, G.J., MACK, S., RUBIN, D.B., SCHAFER, J.L., et ZASLAVSKY, A.M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88, 1149-1166.
- BIEMER, P.P. (1988). Modélisation de l'erreur d'appariement et son effet sur les estimations de l'erreur d'observation du recensement. *Techniques d'enquête*, 14, 125-143.
- BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press.
- CHAPMAN, D.C. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *University of California Publications in Statistics*, 1, 131-160.
- CHILDERS, D., DIFFENDAL, G., HOGAN, H., et MULRY, M. (1989). Coverage Evaluation Research: the 1988 Dress Rehearsal. Communication présentée au Census Advisory Committee of the American Statistical Association et au Census Advisory Committee on Population Statistics à Joint Advisory Committee Meeting, Alexandria, VA.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- JEFFERSON, T. (1986). Lettre au David Humphreys. *The Papers of Thomas Jefferson*, 22, 62.
- JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 test census of Tampa, Florida. *Journal of American Statistical Association*, 84, 414-420.
- JEFFERSON, T. (1986). Lettre au David Humphreys. *The Papers of Thomas Jefferson*, 22, 62.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOGAN, H., et WOLTER, K. (1988). Mesure de l'erreur dans une enquête par téléphone. *Techniques d'enquête*, 14, 105-124.
- HOGAN, H. (1992). The 1990 post-enumeration survey: an overview. *The American Statistician*, 46, 261-269.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of American Statistical Association*, 88, 1047-1060.
- HOG

donne un certain poids à l'affirmation de Fay et coll. lorsque l'appariement pour des données complètes et l'appariement pour des données imputées sont considérées simultanément. Par ailleurs, on peut observer un biais par défaut lorsque β est beaucoup plus grand que γ . Pour la SEP 9, $\beta = .022\%$, soit environ 10 fois la valeur de $\gamma = (.002\%)$. Donc, les fausses concordances prédominent sur les fausses non-concordances pour cette strate, et c'est pourquoi la strate 9 est la seule qui affiche une valeur négative pour Biais(F) ($-.22\%$), ce qui signifie un biais par défaut.

L'application d'une méthode d'appariement suppose toujours un compromis entre les erreurs d'appariement et les cas non résolus. Suivant l'importance du nombre de cas non résolus et l'algorithme d'imputation utilisé, le processus de résolution peut engendrer un nombre appréciable de fausses concordances. Comme nous le disions plus haut, les données empiriques rassemblées par le Bureau of the Census tendent à confirmer le caractère "non biaisé" du mécanisme de non-réponse utilisé dans le processus d'imputation de notre exemple, mais il est souhaitable de recueillir plus de données sur le sujet.

6. CONCLUSION

Dans cet article, nous avons présenté des modèles et des méthodes pour estimer le total d'une population et le sous-dénombrement dans le recensement avec l'intention de compenser le biais d'appariement que renferme l'estimateur de système dual lorsqu'il y a erreur d'appariement. La méthode d'estimation utilisée intègre deux sources d'information: les données de système dual – ou données de saisie-ressaisie – pour le recensement et les données d'une étude sur l'erreur d'appariement (étude de rattachement). La précision de nos estimations repose sur l'hypothèse que le rattachement ne comporte aucune erreur. En outre, il est peu vraisemblable que le taux d'erreur d'appariement soit le même d'une strate à l'autre. Le modèle (B') suppose l'hétérogénéité des taux d'erreur d'appariement dans les diverses strates; toutefois, il exige l'emploi de données de rattachement stratifiées pour estimer les paramètres d'erreur pour les strates. Les méthodes exposées dans cet article sont une généralisation du modèle théorique courant de l'utilisation de l'estimation du maximum de vraisemblance dans le cas où il y a des erreurs d'appariement.

Nous pouvons mesurer l'effet des enrégistrement erronés dans le calcul de EF en nous servant des données d'une étude de rattachement pour l'échantillon D. On obtient le biais d'appariement global de l'ESD en faisant la somme des biais rattachés à l'échantillon P et à l'échantillon D. Notre analyse des données du recensement d'essai effectué à Los Angeles en 1986 indique que l'estimation de système dual du sous-dénombrement dans le recensement renferme un biais par excès d'un peu moins de 1%, ce qui tend à justifier la valeur de 1% utilisée par Hogan et Wolter (1988) dans leur étude d'évaluation. Pour ce qui est de l'analyse des données du recensement de 1990, non seulement les résultats des calculs confirment les caractéristiques déjà connues du biais d'appariement, mais aussi ils révèlent aussi de nouveaux aspects de ce biais.

Comparaison d'estimations basées sur les données pondérées et non-pondérées de l'échantillon P

ANNEXE

Les recherches de M. Fienberg ont été rendues possibles en partie grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada versée à l'Université York de Toronto, Canada. Les auteurs tiennent à exprimer leur reconnaissance à Mary Mulry, pour les données qu'il leur a fournies sur le recensement décennal de 1990, à Joe Sedransk, pour ses suggestions, et à Jay Kadane, Larry Wasserman et Mike Meyer, pour leurs commentaires sur une version antérieure de cet article. Nos remerciements vont aussi au rédacteur associé et aux deux arbitres qui, par leurs commentaires, ont contribué à préciser l'analyse. Les modèles de base utilisés dans cette étude ont été élaborés par le premier auteur lors de la préparation de sa thèse de doctorat à l'Université Carnegie Mellon.

REMERCIEMENTS

Pour des raisons de simplicité, nous avons supposé que l'EF, compte tenu de la stratification, repose sur un échantillonnage aléatoire simple. Toutefois, dans la réalité, les modèles devront être adaptés au degré de complexité du plan d'échantillonnage de l'EF (voir Hogan 1992, 1993). On sait que l'hypothèse de l'appariement parfait ne se vérifie pas quand on applique la méthode d'estimation de système dual dans le contexte du recensement des E.-U. Le problème de l'appariement que pose l'utilisation de l'ESD à deux dimensions. La première concerne l'absence de codes de dénombrement dans l'échantillon P et la seconde a trait aux erreurs qui peuvent être commises lors-qu'il s'agit d'identifier les personnes de l'échantillon P comme des cas recensés ou oubliés. Dans cet article, nous avons présenté une méthode pour résoudre les deux dimensions au moyen de données de système dual corrigées en fonction de probabilités de dénombrement imputées; cette méthode pourrait être utile dans des recensements futurs, pourvu que les modèles soient adaptés au degré de complexité du plan d'échantillonnage de l'EF. Ding (1993c) élabore des estimateurs destinés à résoudre la première dimension du problème de l'appariement en modifiant la méthode d'ESD ordinaire et il décrit le rapport entre les estimations proposées et celles obtenues par suite de l'application de la méthode d'imputation du Bureau of the Census pour les codes de dénombrement manquants de l'échantillon P (Schenker 1988; Belin et coll. 1993).

Pour des raisons de simplicité, nous supposons un poids $k > 1$ pour l'échantillon P et nous considérons la méthode d'estimation de système dual habituelle. Soit $\{x_{ij}\}$ les effectifs par case du tableau 2×2 pour les données pondérées de l'échantillon P et les enrégistrement du recensement, $i, j = 1, 2$ et $ij \neq 22$. On peut connaître par

se produit quand l'ESD (ou l'EMV) est inférieure à CEN, qui est le total des enregistrements du recensement. Les données de système dual sont des chiffres du recensement "redressés" qui, contrairement à CEN, ne comprennent pas les enregistrements erronés.

Les valeurs positives qui figurent dans les colonnes Biais(P), Biais(D) et Biais(T) représentent le biais par excès dont est entachée l'estimation de système dual du sous-dénombrement quand on ne tient pas compte de la source d'erreur correspondante; autrement dit, il faut réduire SD(ESD) d'une valeur équivalente au biais estimé pour tenir compte de l'effet de la source d'erreur. Par ailleurs, nous constatons que dans les quatre cas (SD(ESD), SD(P), SD(D) et SD(T)), les estimations sont beaucoup plus élevées pour les cinq strates de minorités, c'est-à-dire pour SEP 1, 3, 5, 8 et 11. En outre, les estimations des colonnes Biais(P) et Biais(D) sont toutes positives, sauf celle pour la strate 9, dans le cas de Biais(P), et celle pour la strate 11, dans le cas de Biais(D). Ces observations vont dans le sens de l'opinion répandue selon laquelle l'estimation de système dual du sous-dénombrement est normalement entachée d'un biais par excès attribuable aux erreurs d'appariement, sauf dans le cas de certains secteurs géographiques qui ne comptent pas de membres de minorités et où, de fait, les enregistrements erronés sont proportionnellement beaucoup plus nombreux.

Les effets de chaque type d'erreur d'appariement sont donc clairs. Une fausse non-concordance induit un biais par excès, tandis qu'une fausse concordance crée un biais par défaut. La nature du biais d'appariement global dépendra donc de l'importance relative de chaque type d'erreur. Après avoir calculé des estimations du sous-dénombrement pour le recensement de 1980 avec des valeurs choisies de γ et β , Ding (1990) arrive à la conclusion que, étant donné que l'application de la méthode de saisie-ressaisie dans le contexte du recensement suppose des probabilités de saisie élevées, le biais d'appariement est influencé surtout par le taux d'erreur dans les non-concordances (γ) lorsque celui-ci est comparable au taux d'erreur dans les concordances (β). Cette conclusion peut se vérifier facilement dans le contexte de cet article. En effet, la SEP 4 est la strate pour laquelle γ est le plus élevé ($\gamma = .021\%$) et par conséquent celle pour laquelle Biais(P) est le plus élevé (2.06%). Les SEP 3 et 4 présentent à peu près la même estimation de β ($\beta = .012\%$ et $.013\%$ respectivement), mais Biais(P) est beaucoup plus petit pour la strate 3 (.80%) parce que la valeur estimée de γ pour cette strate est plus faible ($\gamma = .010\%$). Donc, une différence d'environ .01% dans la valeur de γ a une influence énorme sur la valeur de Biais(P). Pour les concordances et les non-concordances basées sur des données complètes, Fay et coll. (1988, p. 53) écrivent: "À cause de la nature parfois complexe des opérations d'appariement, les fausses non-concordances représentent peut-être un problème plus grand que les fausses concordances." (TRANSECTION)

Les données que nous avons analysées à l'aide de nos méthodes comprennent aussi bien des données complètes que des données obtenues par imputation au Bureau of the Census. La sensibilité de nos estimations à la valeur de γ

calculées selon le modèle (B'). L'hétérogénéité des probabilités de saisie est appréciable. Cette hétérogénéité et la variabilité des taux d'erreur d'appariement donnent à penser que le modèle (B') convient mieux que le modèle (B). Les écarts-types asymptotiques des tableaux 9 et 11 semblent anormalement faibles par rapport à la taille de l'échantillon de N. Ding (1993b) montre que cette anomalie apparente est une caractéristique propre du système dual lorsque les probabilités de saisie sont très élevées, comme c'est le cas pour un recensement. Malgré des intervalles de confiance très étroits, les études de simulation décrites dans Ding (1993b) montrent que l'approximation normale asymptotique utilisée est très précise du point de vue de la probabilité d'inclusion de la valeur vraie.

Tableau 12
Sous-dénombrement en pourcentage et biais estimés
pour 13 SEP de l'EP de 1990

SEP	SD(ESD)	SD(P)	SD(D)	SD(T)	Biais(P)	Biais(D)	Biais(T)
1*	6.40	5.99	5.30	4.89	0.41	1.10	1.51
2	-0.69	-0.83	-1.05	-1.20	0.14	0.36	0.51
3*	5.59	4.79	5.53	4.72	0.80	0.06	0.87
4	-0.11	-2.17	-1.33	-3.39	2.06	1.23	3.29
5*	5.03	4.49	4.68	4.15	0.53	0.35	0.88
6	1.22	1.06	0.99	0.83	0.16	0.23	0.39
7	1.77	1.73	1.50	1.47	0.03	0.26	0.29
8*	3.52	3.26	3.46	3.20	0.26	0.06	0.32
9	1.05	1.26	1.00	1.21	-0.22	0.05	-0.17
10	0.41	0.34	0.36	0.29	0.07	0.05	0.12
11*	5.26	4.43	5.77	4.94	0.83	-0.51	0.32
12	1.89	1.56	1.51	1.19	0.32	0.38	0.70
13	1.79	1.29	1.28	0.78	0.50	0.51	1.01

Le tableau 12 donne, par rapport à diverses sources, les estimations du biais d'appariement contenu dans les estimations du sous-dénombrement calculées selon la méthode d'ESD ordinaire. SD(ESD) désigne l'estimation de système dual du sous-dénombrement définie comme dans le TOR de 1986; SD(P) est l'estimation du sous-dénombrement calculée au moyen de l'EMV du modèle d'erreur d'appariement et qui prend en compte le biais d'appariement dans l'échantillon P, et Biais(P) = SD(ESD) - SD(P). En outre, d'après Hogg et Tanaka (1988), nous définissons le biais contenu dans l'échantillon D par l'expression Biais(D) = EBR/(ECR + EBR) - EEP/(ECP + EEP) et l'estimation du sous-dénombrement qui prend en compte l'erreur dans l'échantillon D, par SD(D) = SD(ESD) - Biais(D). Enfin, le biais d'appariement total, contenu dans les deux échantillons (P et D), est défini par l'expression Biais(T) = Biais(P) + Biais(D), et l'estimation du sous-dénombrement qui prend en compte les deux sources d'erreur est SD(T) = SD(ESD) - Biais(T). Notons qu'il peut exister des estimations négatives du sous-dénombrement, comme on peut l'observer pour les strates 2 et 4 du tableau 12; cela signifie qu'il y a plutôt surdénombrement. Cette situation

Résultats de l'étude de rattachement pour 13 SEP de l'EP de 1990: échantillon P

SEP	γ_{11}	γ_{21}	γ_{12}	γ_{22}
1*	14,301	124	31	2,773
2	15,051	36	16	1,136
3*	28,784	293	49	4,166
4	32,753	703	27	2,058
5*	28,674	189	18	3,738
6	21,757	69	36	1,156
7	48,061	47	20	3,278
8*	14,800	58	21	2,527
9	16,527	39	20	874
10	43,721	120	107	1,664
11*	12,522	133	11	2,097
12	15,122	59	8	1,078
13	43,356	232	108	4,583

Résultats de l'étude de rattachement pour 13 SEP de l'EP de 1990: échantillon D

SEP	ECP	EFP	ECR	EBR
1*	17,027	1,415	17,106	1,645
2	15,821	879	15,631	932
3*	32,420	2,430	32,322	2,446
4	33,369	1,242	32,922	1,665
5*	32,412	1,880	33,030	2,044
6	24,392	1,225	24,336	1,284
7	51,107	2,908	50,929	3,047
8*	17,174	1,518	17,133	1,526
9	18,279	648	18,228	656
10	44,450	1,604	44,584	1,631
11*	13,644	985	13,693	909
12	15,647	522	15,590	583
13	49,647	2,062	49,545	2,334

Le tableau 9 donne les estimations de système dual – et l'écart-type correspondant – de la probabilité de saisie (qui équivaut à un taux de couverture pour le recensement ou l'échantillon P) pour chacune des treize SEP. Les estimations du tableau 10 indiquent que le taux d'erreur d'appariement varie sensiblement d'une SEP à l'autre. Parmi les trois SEP pour lesquelles γ est plus grand que .01%, la 3 et la 11 sont celles qui rentrent une forte proportion de personnes membres d'une minorité. Il est donc permis de croire que le taux de non-concordances serait plus élevé pour les strates contenant des membres de minorités que pour les autres. En revanche, les estimations de β ne permettent pas d'affirmer que le taux d'erreur dans les concordances est plus élevé, ou moins élevé, pour les strates de minorités. Le tableau 11 donne les estimations du maximum de vraisemblance – et l'écart-type correspondant –

Tableau 10

Estimations du taux d'erreur d'appariement pour 13 SEP de l'EP de 1990

SEP	γ (%)	β (%)
1*	0.009	0.011
2	0.002	0.014
3*	0.010	0.012
4	0.021	0.013
5*	0.007	0.005
6	0.003	0.030
7	0.001	0.006
8*	0.004	0.008
9	0.002	0.022
10	0.003	0.060
11*	0.011	0.005
12	0.004	0.007
13	0.005	0.023

Tableau 11

EMV du modèle (B') et écarts types correspondants pour 13 SEP de l'EP de 1990

SEP	β_1 (E.T.)	β_2 (E.T.)	N (E.T.)
1*	0.92406 (12.68 $\times 10^{-5}$)	0.72114 (18.79 $\times 10^{-5}$)	6,456,833 (446)
2	0.99464 (2.79 $\times 10^{-5}$)	0.93336 (8.30 $\times 10^{-5}$)	9,285,474 (92)
3*	0.93896 (5.38 $\times 10^{-5}$)	0.87597 (7.01 $\times 10^{-5}$)	25,832,352 (279)
4	0.99999 (2.65 $\times 10^{-5}$)	0.98070 (3.64 $\times 10^{-5}$)	30,731,889 (781)
5*	0.94166 (8.28 $\times 10^{-5}$)	0.83080 (12.13 $\times 10^{-5}$)	10,603,717 (306)
6	0.97922 (4.03 $\times 10^{-5}$)	0.95154 (6.03 $\times 10^{-5}$)	14,274,182 (64)
7	0.97600 (2.32 $\times 10^{-5}$)	0.90438 (4.30 $\times 10^{-5}$)	48,717,792 (338)
8*	0.95034 (11.59 $\times 10^{-5}$)	0.86933 (17.06 $\times 10^{-5}$)	4,272,459 (159)
9	0.97756 (4.47 $\times 10^{-5}$)	0.83141 (11.12 $\times 10^{-5}$)	12,097,806 (285)
10	0.99217 (1.50 $\times 10^{-5}$)	0.96733 (3.06 $\times 10^{-5}$)	39,654,306 (90)
11*	0.94239 (10.46 $\times 10^{-5}$)	0.74316 (16.58 $\times 10^{-5}$)	7,729,158 (359)
12	0.97561 (5.07 $\times 10^{-5}$)	0.92614 (8.10 $\times 10^{-5}$)	11,350,674 (101)
13	0.97895 (3.10 $\times 10^{-5}$)	0.99029 (2.42 $\times 10^{-5}$)	26,983,168 (355)

Estimations de système dual et écarts types correspondants pour 13 SEP de l'EP de 1990

SEP	β_1 (E.T.)	β_2 (E.T.)	N (E.T.)
1*	0.92007 (12.57 $\times 10^{-5}$)	0.71803 (18.42 $\times 10^{-5}$)	6,484,855 (470)
2	0.99322 (2.78 $\times 10^{-5}$)	0.93402 (8.17 $\times 10^{-5}$)	9,298,737 (67)
3*	0.93105 (5.33 $\times 10^{-5}$)	0.86858 (6.86 $\times 10^{-5}$)	26,051,987 (540)
4	0.99839 (1.42 $\times 10^{-5}$)	0.96127 (3.46 $\times 10^{-5}$)	31,364,919 (88)
5*	0.93641 (8.22 $\times 10^{-5}$)	0.82618 (11.99 $\times 10^{-5}$)	10,663,134 (390)
6	0.97763 (4.01 $\times 10^{-5}$)	0.95000 (5.83 $\times 10^{-5}$)	14,297,391 (131)
7	0.97567 (2.32 $\times 10^{-5}$)	0.90408 (4.27 $\times 10^{-5}$)	48,734,156 (359)
8*	0.94781 (11.54 $\times 10^{-5}$)	0.86701 (16.85 $\times 10^{-5}$)	4,283,875 (190)
9	0.97699 (4.45 $\times 10^{-5}$)	0.83322 (10.84 $\times 10^{-5}$)	12,071,466 (224)
10	0.99148 (1.48 $\times 10^{-5}$)	0.96665 (2.86 $\times 10^{-5}$)	39,681,946 (108)
11*	0.93419 (10.35 $\times 10^{-5}$)	0.73669 (16.32 $\times 10^{-5}$)	7,797,041 (443)
12	0.97240 (5.05 $\times 10^{-5}$)	0.92309 (8.01 $\times 10^{-5}$)	11,388,243 (164)
13	0.97396 (3.08 $\times 10^{-5}$)	0.98524 (2.35 $\times 10^{-5}$)	27,121,400 (104)

Tableau 9

Nous allons maintenant réanalyser les données du tableau 2 en utilisant le modèle (B), mais nous ne tiendrons pas compte des cas non résolus du tableau 1 parce que nous ignorons à quelle classe ils devraient appartenir. À l'aide des données du tableau 1, nous calculons $\hat{\gamma} = 1 - \hat{\alpha} = .88/(16,623 + 88) = .53\%$ et $\hat{\beta} = 18/(18 + 2,164) = .82\%$. Le tableau 4 donne les estimations – avec l'écart-type correspondant – calculées selon le modèle (B) et selon la méthode d'ESD classique. Les écarts-types sont calculés par la normalité asymptotique; pour plus de détails, voir Ding (1990, 1993a, 1993b). Le sous-dénombrement estimé est alors défini par la formule suivante: sous-dénombrement = $(N - CEN)/N \times 100\%$, où CEN est le total des enregistrés du recensement, c'est-à-dire que CEN = enregistrés corrects du recensement + substitutions + enregistrements erronés (EB) = $343,667 + 5,259 + 6,426 = 355,352$. Les estimations qui figurent sur la dernière ligne du tableau 4 indiquent que l'estimation du sous-dénombrement calculée selon l'ESD devrait être réduite de $.37\%$ (soit 8.42% – 8.05%). Rappelons-nous que Hogan et Wolter (1988) estiment que le taux initial d'EB devrait être relevé de $.5\%$ (soit 2.1% – 1.6%) à cause des données de l'étude de rattachement, ce qui signifie un rajustement additionnel d'environ $.5\%$ du sous-dénombrement estimé. En définitive, nous considérons que l'estimation du sous-dénombrement était entachée d'un biais par excès d'environ $.9\%$ (en supposant que le chevauchement soit négligeable, même si deux éléments ne peuvent, à strictement parler, s'additionner).

Tableau 4

Paramètre	ESD (E.T.)	EMV du modèle (B) (E.T.)
p_1	$.8856 (5.48 \times 10^{-4})$	$.8892 (5.51 \times 10^{-4})$
p_2	$.8677 (5.78 \times 10^{-4})$	$.8712 (5.86 \times 10^{-4})$
N	388,040 (87)	386,470 (79)
Sous-dénombrement (%)	8.42%	8.05%

Comparaison d'estimations pour le recensement d'essai
effectuée à Los Angeles en 1986

5.2 Application du modèle à plusieurs strates au recensement de 1990

Nous allons maintenant analyser des données stratifiées tirées de l'étude d'évaluation de l'EP réalisée dans le cadre du recensement décennal de 1990. Hogan (1993) décrit les opérations et les résultats de l'EP de 1990, Mulry et Spencer (1991, 1993) présentent une analyse de l'erreur totale, et Davis et coll. (1991) parlent de l'étude sur l'erreur d'appariement dans l'EP (Matching Error Study - MES). Cette étude a porté sur treize strates d'évaluation formées à posteriori (SEP) selon la région géographique et le groupe ethnique. Des treize SEP énumérées dans le tableau 5, cinq d'une minorité (Noirs et Latino-Américains): ce sont les

strates 1, 3, 5, 8 et 11. Dans le tableau 6, nous présentons les données de système dual pour chacune des 13 SEP et dans les tableaux 7 et 8, nous donnons les résultats pertinents de l'étude de rattachement pour l'échantillon P et l'échantillon D. Ces résultats sont tirés des rapports finals des projets d'évaluation de l'EP P7 et P10 du Bureau of the Census (Davis et Biemer 1991a, 1991b). L'échantillon P de l'enquête postcensitaire de 1990 comprenait environ 172,000 unités de logement (Hogan 1992). Dans l'analyse habituelle des données de système dual, comme dans l'analyse que nous présentons ici, les données de l'échantillon P sont pondérées pour calculer des estimations de x_{i+1} (total pour l'échantillon P) et de x_{11} (total des concodances). Néanmoins, on peut se servir des données non pondérées (brutes) pour faire de l'inférence; en annexe, nous comparons des estimations basées sur les données réelles de l'échantillon P avec des estimations basées sur les données pondérées du même échantillon.

Tableau 5
13 strates d'évaluation formées a posteriori (SEP)
pour l'EP de 1990

1	Nord-Est, noyau urbain, minorités
2	Nord-Est, noyau urbain, non-minorités
3	Nord-Est, ville non centrale, minorités
4	Nord-Est, ville non centrale, non-minorités
5	Sud, noyau urbain, minorités
6	Sud, noyau urbain, non-minorités
7	Sud, ville non centrale, non-minorités
8	Midwest, noyau urbain, minorités
9	Midwest, noyau urbain, non minorités
10	Midwest, ville non centrale, non-minorités
11	Ouest, noyau urbain, minorités
12	Ouest, noyau urbain, non-minorités
13	Ouest, ville non centrale, non-minorités + Indiens

Tableau 6

Données de système dual pour 13 SEP de 1990

SEP	x_{i+1} (recensement)	x_{i+1} (échantillon P)	x_{i1}
1*	5,966,529	4,656,305.09	4,284,132.78
2	9,235,705	8,685,235.79	8,626,362.34
3*	24,255,611	22,628,349.88	21,068,045.55
4	31,173,378	30,150,266.34	29,966,142.62
5*	9,985,055	8,809,620.02	8,249,407.92
6	13,977,529	13,582,482.34	13,278,614.01
7	47,548,548	44,059,397.93	42,987,517.59
8*	4,060,286	3,714,168.27	3,520,314.04
9	11,826,352	10,058,288.52	9,854,052.95
10	39,343,787	38,358,735.32	38,031,852.01
11*	7,283,885	5,743,998.39	5,365,961.67
12	11,073,872	10,512,339.59	10,222,147.69
13	26,415,232	26,721,116.28	26,025,370.25

* Désigne une strate de minorités formée à posteriori.

Les données recueillies à l'occasion d'une étude de rattachement peuvent être présentées sous la forme illustrée au tableau suivant:

Données d'une étude de rattachement

Classement de rattachement	Classement		Éléments appariés	Éléments non appariés
	Éléments appariés	Éléments non appariés	Éléments appariés	Éléments non appariés
	y_{11}	y_{12}	y_{21}	y_{22}

Pour estimer α et β , nous supposons que dans l'appariement initial les erreurs se produisent selon le modèle (B) et que dans le rattachement elles peuvent ne pas être prises en considération, c'est-à-dire que nous supposons le rattachement parfait. En conséquence, $y_{11} + y_{21}$ est le nombre réel de concordances et il est donc fixe, tandis que y_{11} est une variable aléatoire qui suit une distribution binomiale, c'est-à-dire que $y_{11} \sim \mathcal{B}(y_{11} + y_{21}, \alpha)$. Par conséquent, l'estimation du maximum de vraisemblance de α est $\hat{\alpha} = y_{11} / (y_{11} + y_{21})$ et l'estimation du maximum de vraisemblance du taux d'erreur dans les non-concordances, γ , est $\hat{\gamma} = 1 - \hat{\alpha} = y_{21} / (y_{11} + y_{21})$. Selon le même raisonnement, $y_{12} \sim \mathcal{B}(y_{12} + y_{22}, \beta)$, et l'estimation la plus vraisemblable du taux d'erreur dans les concordances est $\hat{\beta} = y_{12} / (y_{12} + y_{22})$.

Nous pouvons nous servir des estimations du taux d'erreur d'appariement calculées ici pour analyser les données de l'étude de rattachement faite lors du recensement d'essai de Los Angeles. Très souvent, il est intéressant d'estimer, outre la taille d'une population, celle d'une sous-population définie selon des critères démographiques (par ex., Noirs, Blancs) ou des critères géographiques. Dans ce cas, il convient mieux d'imaginer des taux d'erreur d'appariement différents d'une strate à l'autre en utilisant des estimations du taux d'erreur d'appariement pour chaque strate à l'étude. On peut obtenir des estimations de ce genre en effectuant une étude de rattachement pour chaque strate, puis en se servant des estimations tirées de cette étude. Les données servant à l'application du modèle (B') sont tirées de la base du recensement de 1990, et nous les analysons ici.

5. APPLICATIONS

5.1 Application du modèle à une strate au TOR de 1986

Hogan et Wolter (1988) présentent les données de l'étude de rattachement effectuée à l'occasion du TOR de 1986 à Los Angeles. Les résultats de l'étude pour l'échantillon P sont exposés dans le tableau 1 sous la forme d'une table où se recoupent les codes d'appariement obtenus par suite de l'appariement initial du TOR et ceux obtenus par suite du nouvel appariement. Le tableau 2 présente, sous la forme d'une table à double entrée, les données pour le TOR de

1986, sans stratification a posteriori. Le nombre estimé de personnes qui n'ont été recensées par aucun des deux systèmes (5,870) est du même ordre de grandeur que le nombre de substitutions dans le recensement (5,259) et que le nombre d'enregistrements erronés (6,426) (Hogan et Wolter 1988). Les résultats de l'étude pour l'échantillon D figurent dans le tableau 3. Posons ECP et EEP comme le nombre total d'enregistrements corrects et que le nombre total d'enregistrements erronés, respectivement, selon le classement initial, et posons ECR et EBR comme le nombre total d'enregistrements corrects et le nombre total d'enregistrements erronés, respectivement, selon le rattachement. Se fondant sur les données du tableau 3, Hogan et Wolter (1988) concluent que le taux initial d'enregistrements erronés (EB) - EEP/(ECP + EEP) = 325/(325 + 19,269) = .016 - devrait être porté à environ .021, c'est-à-dire que EBR/(ECR + EBR) = 411/(411 + 19,334).

Tableau 1

Résultats de l'étude de rattachement effectuée dans le cadre du recensement d'essai effectué à Los Angeles en 1986: échantillon P. Source: Hogan et Wolter (1988)

Classement de rattachement		Classement du rattachement	
l'appariement initial	Appariés	Non appariés	Total
		Non appariés	Total
Appariés	16,623	18	55
Non appariés	88	2,164	56
Non résolus	17	0	132
Total	16,728	2,182	243

Tableau 2

Données et estimations de système dual pour le recensement d'essai effectué à Los Angeles en 1986 Source: Hogan et Wolter (1988)

ÉP		Dénombrés Manqués Total	
Enregistrements corrects du recensement*	Manqués	Dénombrés	Total
		Manqués	Total
343,667	45,463	298,204	388,040
44,373	5,870	38,503	51,333
Total		336,707	388,040

* Enregistrements corrects du recensement = total des enregistrements du recensement - substitutions - enregistrements erronés.

Tableau 3

Résultats de l'étude de rattachement effectuée dans le cadre du recensement d'essai effectué à Los Angeles en 1986: échantillon D. Source: Hogan et Wolter (1988)

Résultats initiaux		Résultats du rattachement	
Enregistrements corrects	Enregistrements erronés	Enregistrements corrects	Total
		Enregistrements erronés	Total
19,153	28	283	19,269
41	100	1	463
19,334	411	223	20,057
Total		312	20,057

Nous utilisons la méthode de vraisemblance conditionnelle élaborée par Sanathanan (1972). Pour une valeur n fixe, (x_{11}, x_{12}, x_{21}) a une distribution multinomiale avec une fonction de vraisemblance

$$L_1(p_{11}, p_{12}, p_{21}) = \frac{n!}{x_{11}! x_{12}! x_{21}!} \cdot \frac{(p_{11} + p_{12} + p_{21})^n}{p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}}} \quad (1)$$

On considère alors que n suit une distribution binomiale avec une taille d'échantillon N et une probabilité $p_{11} + p_{12} + p_{21}$, et la fonction de vraisemblance correspondante est

$$L_2(N) = \frac{N!}{n!} \frac{(1 - (p_{11} + p_{12} + p_{21}))^{N-n}}{(p_{11} + p_{12} + p_{21})^n} \quad (2)$$

Selon l'approche conditionnelle, nous calculons les estimations du maximum de vraisemblance des probabilités par case au moyen de l'équation (1), puis, à l'aide des valeurs des probabilités par case, nous déterminons la valeur de N qui maximise (2). Sanathanan (1972) a montré que dans des conditions de régularité convenables, les estimations de vraisemblance conditionnelle aussi bien que les estimations de vraisemblance inconditionnelle de N sont convergentes et ont la même distribution normale multivariée asymptotique. L'approche conditionnelle convient particulièrement bien aux situations qui impliquent de gros échantillons, comme c'est le cas ici.

Suivant l'hypothèse de l'unité formite des probabilités de sélection, nous posons p_1 comme la probabilité qu'une unité quelconque de la population soit incluse dans X_1 et, de la même manière, nous posons p_2 comme la probabilité qu'une unité de la population soit incluse dans X_2 . On appelle habituellement ces probabilités p_1 et p_2 des probabilités de saisie, et elles ne dépendent pas de la manière dont s'opère l'appariement. Donc, la probabilité qu'un individu se trouve dans les deux échantillons est $p_1 p_2$ et la probabilité de faire partie de l'ensemble N_1 est $p_1 (1 - p_2)$. Puisque le modèle (A) est un cas particulier du modèle (B), avec $\beta = 0$, nous allons nous attarder à formuler le problème dans les termes du modèle (B). Pour cela, il faut d'abord établir la spécification paramétrique des probabilités par case. Deux scénarios seulement feront qu'un individu pourra se trouver dans la case (1, 1) du tableau de contingence: soit que l'individu appartienne effectivement aux deux échantillons, et un appariement est alors effectué, ou bien l'individu appartient à N_1 et il est apparié erronément à un individu de N_2 pour employer la notation précédente dans la dernière section. Dans ce cas-ci, la direction de l'appariement, de N_1 vers N_2 , est implicite dans l'hypothèse (iii) du modèle (B). La probabilité de réalisation du premier scénario est $\alpha p_1 p_2$ et la probabilité de réalisation du second est $\beta p_1 (1 - p_2)$. En outre, les deux scénarios s'excluent mutuellement. Par conséquent, nous avons $p_{11} = \alpha p_1 p_2 + \beta p_1 (1 - p_2)$ et $p_{12} = p_1 - p_{11} = p_1 -$

4. ESTIMATION DES TAUX D'ERREUR D'APPARIEMENT À L'AIDE DES DONNÉES D'UNE ÉTUDE DE RAPPARIEMENT

Dans cette section, nous calculons des valeurs estimées des paramètres α et β du taux d'erreur d'appariement en nous servant des données d'une étude sur l'erreur d'appariement (étude de rappariement) qui s'inscrivait dans le recensement d'essai qu'a effectué en 1986 sur le territoire de Los Angeles le Bureau of the Census dans le but d'évaluer l'EB. En quelques mots, une étude de rappariement porte ordinairement sur un échantillon de cas; sa réalisation exige l'application de méthodes plus poussées, l'emploi de personnel hautement qualifié et le recours à la réinterview pour obtenir des estimations du biais lié aux opérations d'appariement antérieures. Pour plus de détails, voir Childers, Diffendal, Hogan et Mulry (1989). Dans leur analyse de l'étude de rappariement effectuée à l'occasion du TOR de Los Angeles, Hogan et Wolter (1988) écrivent: "On a effectué [le rappariement] indépendamment de l'appariement initial, puis on a déterminé quels étaient les écarts entre les résultats des deux opérations. À cause de la rigueur avec laquelle s'est fait le nouvel appariement, nous croyons que les résultats de cet appariement reflètent la réalité tandis que les écarts entre les résultats du premier et du second appariement reflètent le biais dont sont entachés les résultats du premier appariement."

$$N = \frac{p_1 + p_2 - (\alpha - \beta) p_1 p_2 - \beta p_1}{n} \quad (3)$$

(voir Chapman 1951). En ce qui concerne les modèles (A') ou (B'), on peut se servir, pour la strate i , des estimations de paramètres calculées selon le modèle (A) ou (B) pour les données de cette strate, puis faire la somme des estimations pour toutes les strates afin d'estimer le total de population.

exprimé par la formule

Si $\alpha = 1$ et $\beta = 0$, la situation ci-dessus se ramène au cas classique des deux échantillons, et il existe alors des solutions complètes bien connues pour les équations de vraisemblance liées à la fonction de vraisemblance conditionnelle (1) (voir Bishop et coll. 1975, chap. 6, p. 232), ce qui nous amène à l'estimateur de système dual ordinaire, $N_{ESD} = x_{1+} x_{+1} / x_{11}$. Autrement, il n'existe pas d'expression en forme analytique fermée pour les estimations du maximum de vraisemblance. Cependant, lorsqu'on connaît p_1 et p_2 , qui sont les estimations du maximum de vraisemblance conditionnelle de p_1 et p_2 , l'estimateur du maximum de vraisemblance conditionnelle de N peut être

Ses conditions sont satisfaites par le paramétrage de $\{p_{ij}\}$ dans le cas ci-dessus.

Nous remarquons que les cas a), b) ou c) peuvent se produire seulement si au moins deux erreurs sont commises: le non-appariement de deux unités appariables et l'appariement de deux unités qui n'ont pas de lien entre elles. Comme la probabilité de ces erreurs est faible, nous supposons, pour des raisons de simplicité, que les cas a), b) et c) ont très peu de chances de se produire selon le modèle suivant.

Modèle (B):

- i) comme dans le modèle (A), les concordances réalisées entre M_1 et M_2 seront de nouveau réalisées, mais avec une probabilité α , $0 < \alpha \leq 1$;
- ii) les fausses concordances de type a), b) ou c) ont une probabilité de réalisation négligeable;
- iii) chaque unité (ou individu) de N_1 sera appariée à une unité de N_2 avec une probabilité commune β , $0 \leq \beta < 1$.

Même si, en théorie, la valeur de α et celle de β peuvent varier de 0 à 1, dans le contexte du recensement nous nous attendons que $\alpha \approx 1$ et $\beta \approx 0$.

Nous pouvons aussi envisager le cas où les probabilités d'erreur d'appariement et les probabilités de saisie sont plus homogènes dans ces strates que dans la population totale. Supposons que la population totale consiste en l strates. Soit $Z_{N_i \times l}$ le vecteur de caractéristiques pour la population de la strate i de taille inconnue N_i , et soient Y_{i1} , Y_{i2} deux échantillons prélevés dans la strate i qui servent à calculer N_i . Nous pouvons alors construire un estimateur de la taille de la population globale en posant $N = \sum_{i=1}^l N_i$. Nous pouvons affiner les modèles (A) et (B) de la façon suivante:

Modèle (A):

Le modèle (A) est valide pour chaque strate et α_i est la probabilité d'appariement de deux éléments appariables de la strate i , $0 < \alpha_i \leq 1$, $1 \leq i \leq l$.

Modèle (B):

Le modèle (B) est valide pour chaque strate et α_i , β_i , $1 \leq i \leq l$, sont les deux paramètres de probabilité pour la strate i .

Dans l'enquête postcensitaire de 1990, l'étude d'appariement pour l'échantillon P s'est faite au moyen des ilots échantillonnés et d'un anneau d'ilots avoisinants (Hogan 1993). Les erreurs de géocodage peuvent amener l'établissement de fausses concordances entre des éléments de strates formées à posteriori selon des critères géographiques; de fausses concordances sont également possibles pour les strates définies selon des critères démographiques. Le modèle (B') suppose implicitement qu'il ne peut y avoir de fausse concordance entre les éléments de strates formées à posteriori. En outre, tous les modèles sont une simplification du plan de sondage de l'EP.

3. ESTIMATION DU TOTAL D'UNE POPULATION

Dans cette section, nous cherchons à estimer le total d'une population suivant les divers modèles d'erreur d'appariement - (A), (A'), (B) et (B') - en supposant valides les hypothèses habituelles de l'indépendance des deux échantillons et de l'uniformité des probabilités de sélection. Pour des modèles où l'on suppose que les probabilités d'échantillonnage ne sont pas uniformes ou que les échantillons ne sont pas indépendants, voir la méthode à trois échantillons dans Darroch et coll. (1993) et la méthode décrite dans Alho et coll. (1993).

Soit N le nombre d'unités (ou d'individus) formant la population à l'étude, x_{1+} , le nombre d'unités dans Y_1 , x_{+1} , le nombre d'unités dans Y_2 , et x_{11} , le nombre d'unités présentes dans les deux échantillons à la fois. Le nombre d'unités présentes dans Y_2 mais non dans Y_1 est $x_{21} = x_{+1} - x_{11}$ et le nombre d'unités présentes dans Y_1 mais non dans Y_2 est $x_{12} = x_{1+} - x_{11}$. On peut disposer les données de saisie-ressaisie dans un tableau de contingence 2×2 dont une des cases est vide:

Echantillon Y_2		unités présentes	unités absentes
unités présentes	unités absentes	x_{11}	x_{12}
unités présentes	unités absentes	x_{21}	—

Dans ce tableau on utilise le symbole "—" pour désigner une case vide, ainsi que la notation habituelle pour les totaux marginaux: $x_{1+} = x_{11} + x_{12}$, $x_{+1} = x_{11} + x_{21}$. À ce tableau correspond un tableau de probabilités - $p_{ij} = \text{Pr}$ [un individu se trouve dans la case (i,j)] - de même dimension,

Echantillon Y_2		unités présentes	unités absentes
unités présentes	unités absentes	p_{11}	p_{12}
unités présentes	unités absentes	p_{21}	p_{22}

étant donné la contrainte linéaire habituelle

$$\sum_{j=1}^2 \sum_{i=1}^2 p_{ij} = 1.$$

Soit n le nombre d'individus différents présents dans les deux échantillons, c'est-à-dire que $n = x_{11} + x_{12} + x_{21}$. Si nous supposons que les échantillons sont tirés aléatoirement avec probabilités de sélection uniformes pour les unités, les effectifs par case ont une distribution multinomiale

$$(x_{11}, x_{12}, x_{21}, N - n) \sim \text{Mult}(N, p_{11}, p_{12}, p_{21}, p_{22}).$$

L'appariement sert à déterminer un code de dénombrement pour les personnes qui forment l'échantillon P. Plus précisément, les membres de l'échantillon P qui sont apparus à un enrégistrement du recensement sont considérés comme ayant été dénombrés. Les membres de l'échantillon P qui ne peuvent être apparus à aucun enrégistrement du recensement sont, pour la plupart, considérés comme ayant été oubliés. Deux raisons générales expliquent que des erreurs d'appariement puissent se produire:

1. soit que l'information enrégistrée par le répondant ou l'intervieweur est inexacte;

2. soit que l'information enrégistrée est exacte mais utilisée incorrectement.

En outre, deux types d'erreur sont possibles: l'appariement de cas non appariables et le non appariement de cas appariables. Le premier type d'erreur peut être divisé en deux catégories:

a) le cas où un membre de l'échantillon P a été apparié par erreur à l'enrégistrement d'une autre personne, alors qu'il aurait dû être apparié à un enrégistrement de l'échantillon D;

b) le cas où aucun appariement n'aurait dû être fait.

La première catégorie n'a pas de conséquences "graves" pour l'estimation de N puisque en fait la personne "mal appariée" aurait normalement été comptée parmi les enrégistres du recensement. En revanche, les erreurs de la seconde catégorie ont pour effet de provoquer une sous-estimation du nombre de non-concordances. Par ailleurs, le deuxième type d'erreur (non appariement de cas appariables) a pour effet de surestimer le nombre de non-concordances. Fay, Passel, Robinson et Cowan (1988) soulignent que les fausses non-concordances représentent sans doute un problème plus grand que les fausses concordances. Celles-ci sont plus rares que celles-là parce qu'il est facile de vérifier l'authenticité d'une concordance.

Dans la section 2, nous proposons des modèles d'erreur d'appariement et dans les sections 3 et 4, nous présentons une méthode systématique pour estimer le total d'une population et, partant, le sous-dénombrément dans le recensement. Dans la section 5, nous analysons les données du recensement d'essai de 1986 effectué à Los Angeles et du recensement décennal de 1990 afin de montrer comment notre méthode évalue l'erreur d'appariement contenue dans les estimations du sous-dénombrément.

2. MODÉLISATION DE L'ERREUR D'APPARIEMENT

Pour des raisons de simplicité, nous appliquons une contrainte au mécanisme d'appariement, à savoir qu'un membre d'un échantillon ne peut être apparié à plus d'un membre d'un autre échantillon. En outre, nous supposons implicitement un échantillonnage aléatoire stratifié, ce qui donne un modèle d'échantillonnage multinomial classique pour l'estimation de système dual. Cette simplification nous permet de concentrer notre attention sur l'effet de

L'objectif de l'étude est d'établir des modèles d'appariement dans le but explicite d'établir des modèles d'appariement. Soit $Z^{N \times 1}$ le vecteur de caractéristiques pour la population entière, de sorte que la i -ième composante de $Z^{N \times 1}$ est $1 \leq i \leq N$. Ce ne sont pas toutes les composantes de $Z^{N \times 1}$ qui peuvent être observées dans un échantillon quelconque. Nous cherchons à estimer N, la taille de la population, à partir des données de deux échantillons. Imaginons que l'action de tirer un échantillon dans une population équivaut à prélever aléatoirement des composantes de $Z^{N \times 1}$ pour former un nouveau vecteur Y. Si les composantes prélevées ne contiennent pas toutes les caractéristiques voulues ou si elles renferment des caractéristiques erronées, il peut se produire des erreurs d'appariement. Nous allons donc désigner le premier échantillon par X_1 et le second, par X_2 , et, dans l'analyse qui va suivre, les deux serviront d'échantillon de saisie-ressaisie pour l'estimation de système dual.

Deux types d'erreur d'appariement sont possibles: le non-appariement de cas appariables (que nous appellerons erreur de type 1) et l'appariement de cas non appariables (que nous appellerons erreur de type 2). Nous pouvons choisir de modéliser l'un ou l'autre type d'erreur ou les deux. Si l'appariement est parfait, chaque composante de X_1 ou X_2 contiendra la même information que dans $Z^{N \times 1}$, et le nombre de concordances sera égal au nombre d'éléments communs à X_1 et X_2 . Si l'appariement est imparfait, nous considérerons le modèle élémentaire suivant:

Modèle (A):

i) les concordances réalisées dans les conditions d'appariement parfait seront de nouveau réalisées, avec une probabilité commune α , $0 < \alpha \leq 1$;

ii) toutes les non-concordances demeureront des non-concordances, c'est-à-dire qu'il n'y aura pas de fausses concordances.

Le modèle (A) définit le mécanisme de l'erreur d'appariement de type 1 avec probabilité d'erreur $1 - \alpha$, en supposant que l'erreur de type 2 soit négligeable.

Pour élaborer un modèle pour les deux types d'erreur d'appariement, nous devons examiner soigneusement toutes les situations qui peuvent mener à de fausses concordances. Lorsqu'il n'y a pas d'erreur d'appariement, on peut écrire $X_1 = (M_1, N_1)$ et $X_2 = (M_2, N_2)$, de sorte que les ensembles M_1 et M_2 ont la même taille et que chaque unité de M_1 est appariée correctement à une unité de M_2 et vice versa, que N_1 est l'ensemble des unités de l'échantillon X_1 qui ne sont appariées à aucune des unités de X_2 qui ne sont appariées à aucune des unités de X_1 . Lorsque des erreurs d'appariement sont possibles, de fausses concordances peuvent être établies selon quatre possibilités:

a) une unité de M_1 est appariée incorrectement à une unité de M_2 ;

b) une fausse concordance est établie entre M_1 et N_2 ;

c) une fausse concordance est établie entre M_2 et N_1 ;

d) une fausse concordance est établie entre N_1 et N_2 .

Estimation de système dual du sous-dénombrement recensement lorsqu'il y a erreur d'appariement

YE DING et STEPHEN E. FIENBERG¹

RÉSUMÉ

Depuis 1950, le Bureau of the Census des E.-U. a recouru à l'estimation de système dual (ESD) pour mesurer le taux de couverture du recensement décennal. Selon cette méthode, on combine des données d'un échantillon avec des données du recensement pour estimer le sous-dénombrement et le surdénombrement dans le recensement. L'ESD repose sur l'hypothèse d'un appariement parfait des unités du recensement et de celles de l'échantillon. Or, les erreurs d'appariement et les fausses non-concordances, qui sont inévitables, ont pour effet de réduire la précision de cette méthode d'estimation. Dans cet article, nous reconsidérons l'ESD en assouplissant l'hypothèse de l'appariement parfait et nous proposons des modèles qui décrivent deux types d'erreur d'appariement: l'appariement de cas qui ne sont pas appariables et le non-appariement de cas qui sont appariables. En outre, nous présentons des méthodes pour estimer le total d'une population et le sous-dénombrement dans le recensement et nous en illustrons l'application au moyen de données du recensement d'essai de 1986 à Los Angeles et du recensement décennal de 1990.

MOTS CLÉS: Saisie-ressaisie; biais d'appariement; modélisation de l'erreur d'appariement; fonction de vraisemblance multinomiale.

1. INTRODUCTION

Aux États-Unis, le problème du sous-dénombrement dans le recensement est un sujet majeur de préoccupation depuis le tout premier recensement, effectué en 1790 (Jefferson 1986). Dans le cadre de ce qu'on appelle le programme d'enquêtes postcensitaires (EP), on utilise l'estimation de système dual (ESD) (ou méthode de saisie-ressaisie) pour évaluer le taux de couverture de la population dans le recensement. Erickson et Kadan (1985) et Wolter (1986) décrivent l'utilisation de l'ESD par rapport au recensement décennal de 1980. En ce qui concerne le recensement décennal de 1990, un nouveau plan de sondage était prévu pour l'EP et on a cherché à perfectionner la méthodologie à l'occasion d'un recensement d'essai effectué en 1986 dans le Central Los Angeles County et connu sous le nom de Test des opérations de redressement (TOR). Diffendal (1988) traite de la méthodologie, des opérations et des résultats du TOR, tandis que Hogan et Wolter (1988) et Schenker (1988) font une évaluation des opérations et des hypothèses qui se rattachent à l'ESD.

Dans l'EP, l'estimation de système dual emploie deux échantillons: l'échantillon P et l'échantillon D. Le premier, tout à fait indépendant du recensement, sert à mesurer les omissions du recensement; le second, formé d'enregistrements du recensement, sert à mesurer les enregistrements erronés. Pour le TOR de 1986, l'estimateur de système dual de la taille de la population, N , qui combine l'information de l'échantillon P et celle de l'échantillon D, était défini par l'expression suivante:

$$N = (CEN - EE - SUB) \cdot N_p/M,$$

où CEN est le chiffre du recensement non redressé, EE, le nombre estimé d'enregistrements erronés et de personnes non appariables dans le recensement, SUB, le nombre de substitutions de personnes dans le recensement, N_p , le nombre de personnes dans l'échantillon P, et M , le nombre estimé de personnes qui font partie à la fois de la population recensée et de l'échantillon P. Pour plus de détails, voir Diffendal (1988) ou Wolter (1986). Une variante de cette formule a été utilisée dans le cadre du recensement de 1990; voir à ce sujet Hogan (1992, 1993).

L'ESD et le problème de l'appariement ont fait l'objet d'une grande attention dans les années 1970 à cause de l'emploi de cette méthode pour l'estimation du nombre de naissances et de décès dans les pays en voie de développement, et certains croient que le taux d'erreur d'appariement a peut-être été la principale difficulté de la méthode d'estimation de système dual utilisée dans le recensement de 1980 (Fienberg 1989). Jaro (1989) décrit les nouvelles techniques d'appariement introduites par le Bureau of the Census pour 1990 ainsi que l'expérimentation qui a été faite de la méthodologie correspondante lors d'un prétest en 1985. Bieher (1988) examine des modèles servant à mesurer l'effet de l'erreur d'appariement sur les estimations de l'erreur de couverture du recensement, sans tenter de compenser le biais d'appariement contenu dans l'estimation de système dual ordinaire. La méthode utilisée dans le recensement de 1990 comprenait non seulement un algorithme d'appariement informatisé et des étapes de suivi manuel, mais aussi des modèles de régression logistique pour les cas non résolus des échantillons P et D (voir Belin et coll. 1993).

¹ Ye Ding est Research Scientist, Bureau of Biometrics, New York State Health Department, Concourse, Room C-144, Empire State Plaza, Albany, New York 12237, U.S.A.; Stephen E. Fienberg est Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, U.S.A.

6. CONCLUSIONS

On sait depuis longtemps que, sauf dans de rares cas, la non-réponse entraîne un biais dans les estimations d'une enquête. L'imputation est une méthode utilisée couramment pour traiter la non-réponse, parce qu'il est commode de travailler avec un ensemble de données complet. Pour les grandes enquêtes, on peut employer de nombreuses règles d'imputation ainsi que certains logiciels. On applique parfois l'imputation sans examen critique et, bien qu'elle soit utilisée couramment, l'imputation ne résout pas le sérieux problème du biais causé par la non-réponse.

Dans le présent article, nous avons étudié l'imputation par quotient. L'imputation par quotient ordinaire $B_{j \cdot x_k}$ est justifiée (c.-à-d. qu'elle ne produit pas de biais) si deux conditions sont respectées: (a) le modèle de régression qui sous-jacent la règle d'imputation par quotient s'applique (c.-à-d. qu'il s'agit d'une régression linéaire passant par l'origine); (b) le mécanisme de réponse est non-confondu. Les résultats de notre simulation donnent une certaine idée de la valeur de l'estimateur d'imputation par quotient ordinaire \hat{y}^{raimp} quand l'une de ces conditions ou les deux ne sont pas respectées. Nous avons étudié plusieurs mécanismes de réponse non uniformes, confondu ainsi que non-confondu. Nous avons aussi étudié les cas où le modèle de régression qui est à la base de l'imputation par quotient ne s'applique pas.

Nous avons soutenu qu'on peut parfois supposer de façon réaliste l'existence d'un mécanisme confondu dans une enquête. Nous avons montré que si une hypothèse de mécanisme de réponse confondu est justifiée et si le modèle qui sous-jacent l'imputation par quotient est valable, on peut obtenir une certaine réduction du biais par les estimateurs corrigés en s présentés dans l'article. Ces estimateurs ont un biais beaucoup moins élevé que l'estimateur non corrigé \hat{y}^{raimp} . Ils sont aussi généralement plus efficaces que les estimateurs corrigés en r sur ce plan.

Supposons que l'analyste travaille avec l'hypothèse que le modèle d'imputation par quotient (2.2) s'applique. Notre étude de simulation nous amène alors à proposer des estimateurs selon la nature supposée du mécanisme de réponse et le taux de non-réponse (tableau 5). L'inscription "tous" signifie que l'on peut utiliser l'un ou l'autre des 10 estimateurs présentés au tableau 2.

Tableau 5

Estimateurs proposés pour chaque mécanisme de non-réponse

Taux de non-réponse	Estimateur proposé		
	Mécanisme de réponse	Uniforme	Non-confondu
(≤ 10%)	tous	tous sauf y_r	tous sauf y_r
(> 10%)	tous ¹	\hat{y}^{raimp} corrigé en s	

Note 1: \hat{y}^{raimp} possède un léger avantage par rapport aux autres estimateurs.

BIBLIOGRAPHIE

BAKER, S.G., et LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 83, 62-69.

FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

FAY, R.E. (1989). Estimating nonignorable nonresponse in longitudinal surveys through causal modeling. Dans *Panel Surveys* (Eds. D. Kasprzyk, G.J. Duncan, G. Kalton, et M.P. Singh), 375-399.

GREENLESS, J.S., REECE, W.S., et ZIESCHANG, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 82, 251-261.

LEE, H., RANCOURT, E., et SÄRNDAAL, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

RUBIN, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72, 538-543.

RUBIN, D.B. (1986). Initiation à l'imputation multiple pour les cas de non-réponse. *Techniques d'enquête*, 12, 41-52.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

SÄRNDAAL, C.-E. (1990). Méthodes pour estimer la précision des estimations d'enquête lorsqu'il y a eu imputation. *Recueil: Symposium 90, Mesure et amélioration de la qualité des données*, Statistique Canada, 369-380.

SÄRNDAAL, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

7. REMERCIEMENTS

Les auteurs désirent remercier les arbitres ainsi que le rédacteur associé pour leurs précieux commentaires. Une version antérieure à cet article a été présentée à la Conférence annuelle de recherche à Arlington en Virginie, du 22 au 25 mars 1992.

Si le modèle de régression qui est à la base de l'imputation par quotient ne s'applique pas, la situation est moins claire. À moins que l'hypothèse simpliste d'un mécanisme de réponse uniforme ne s'applique (ce qui est peu probable), l'estimateur d'imputation par quotient non corrigé \hat{y}^{raimp} peut être entaché d'un biais considérable. Nous avons trouvé que \hat{y}^{raimp} est particulièrement sujet au biais pour une population de type CONVEXE, pour laquelle les estimateurs corrigés en s ont habituellement un biais inférieur à \hat{y}^{raimp} . Par contre, pour les populations de type CONCAVE et NON-QUOTIENT, \hat{y}^{raimp} résiste généralement mieux au biais que les estimateurs corrigés en s.

(iv) Le mécanisme (M5) (confondu et dépendant des valeurs y).

Le tableau 2 montre que pour les estimateurs corrigés en s , la VABR est plus faible que pour l'estimateur non corrigé y^{raimp} ; la RCEQMR est légèrement plus élevée toutefois pour ces estimateurs. Par contre, les estimateurs corrigés en r "vont trop loin", de sorte que la VABR et la RCEQMR dépassent les niveaux observés pour y^{raimp} . Le groupe des estimateurs corrigés en r ne donne pas de bons résultats pour ce mécanisme.

En résumé, le tableau 2 montre que si le modèle d'imputation par quotient (2.2) s'applique et si l'hypothèse d'un mécanisme confondu est juste, la décision d'utiliser un des estimateurs corrigés peut entraîner une réduction du biais. La principale difficulté que rencontre l'analyste est celle de prédire avec exactitude la nature du mécanisme de réponse qui cause la non-réponse. En particulier, il se peut que l'analyste ait de la difficulté à faire la distinction entre un mécanisme confondu (p. ex., avec $\Theta_k = e^{-\gamma_k}$) et un mécanisme semblable non-confondu et non uniforme (p. ex., avec $\Theta_k = e^{-\gamma_{Xk}}$). Cependant, cette différence, aussi subtile soit-elle, a une incidence considérable sur le biais de y^{raimp} et sur la décision d'utiliser ou non un estimateur corrigé. Quand il s'agit d'un mécanisme non-confondu non uniforme, nous avons vu que des conséquences négatives se rattachent aux estimateurs corrigés, en particulier dans le cas des estimateurs corrigés en r .

5.2 Autres types de régression

Le tableau 3 présente le comportement de six estimateurs (les deux estimateurs non corrigés et les quatre estimateurs corrigés en s) pour les types de régression CONCAVE, CONVEXE et NON-QUOTIENT. Comme dans le tableau 2, il existe peu de différence entre les estimateurs quand le mécanisme uniforme (M1) s'applique. Pour les deux mécanismes confondu, il ne ressort pas clairement des résultats présentés au tableau 3 qu'il faudrait effectuer une estimation corrigée en s , même si l'hypothèse d'un mécanisme confondu est faite à bon droit. Comparativement à l'estimateur non corrigé y^{raimp} , les estimateurs corrigés en s montrent une amélioration nette

Tableau 4

VABR (BR), RCEQMR (MR) et NCR moyens des deux estimateurs non corrigés et des

estimateurs corrigés par c_q et k_q

(Moyenne calculée sur tous les types de population)

	M1			M2			M3			M4			M5			Globalement		
	BR	MR	NCR	BR	MR	NCR	BR	MR	NCR	BR	MR	NCR	BR	MR	NCR			
y_r	0.3	14.7	92.3	13.0	20.0	86.8	9.8	17.1	80.2	19.1	24.4	77.0	17.4	22.1	65.6	11.9	19.6	80.4
y_r^{raimp}	0.3	13.2	93.1	2.5	13.0	92.9	3.1	14.0	92.0	5.8	14.2	93.0	9.3	16.7	81.0	4.2	14.2	90.4
$y_r^{cd \cdot s}$	1.0	14.2	92.3	4.7	14.0	86.8	8.3	19.0	90.8	3.7	13.2	91.5	8.1	17.7	87.0	5.2	15.6	89.7
$y_r^{k4 \cdot s}$	1.1	14.0	92.9	3.8	13.6	89.5	7.4	18.4	92.2	3.6	13.3	92.6	6.7	17.0	88.0	4.5	15.2	91.0

5.4 Commentaires généraux

L'étude du tableau 4 (un tableau récapitulatif) nous permet de constater que, comme prévu, c'est y_r et y^{raimp} qui ont le meilleur comportement pour le mécanisme de réponse uniforme (M1). L'estimateur non corrigé y^{raimp} est le meilleur pour les mécanismes non-confondu (M2) et (M3), alors que les estimateurs corrigés sont les meilleurs pour les mécanismes confondu (M4) et (M5).

Finalement, en moyenne, pour l'ensemble des 240 cas inclus dans notre étude, nous remarquons d'après les données qui figurent dans la colonne "Globalement" du tableau 4 que y^{raimp} et $y^{k4 \cdot s}$ ont un comportement semblable, le premier ayant un biais légèrement plus faible et le second, un niveau de confiance réel légèrement supérieur.

Tableau 3

VABR (BR), RCEQMR (MR) et NCR moyens de six estimateurs pour des populations de type CONCAVE, CONVEXE et NON-QUOTIENT

(Pour chaque mécanisme, la moyenne de 12 cas a été calculée comme pour le tableau 2)

	M1			M2			M3			M4			M5		
	BR	MR	NCR	BR	MR	NCR	BR	MR	NCR	BR	MR	NCR	BR	MR	NCR
\bar{y}_r	0.2	10.4	92.9	10.5	14.8	82.3	7.3	12.7	82.3	12.3	16.0	78.3	8.7	13.4	78.8
$\bar{y}_{r\text{raim}}$	0.2	9.4	94.5	1.4	9.1	93.4	2.6	10.5	94.9	1.9	9.2	94.9	2.1	9.7	92.9
$\bar{y}_{c2.s}$	1.1	11.4	92.4	6.3	11.4	84.7	11.8	18.8	88.4	3.2	10.2	90.0	5.5	14.2	92.3
$\bar{y}_{c4.s}$	1.0	11.1	92.8	6.6	11.5	84.3	11.4	18.0	88.8	3.6	10.3	89.8	5.5	13.7	92.7
$\bar{y}_{k2.s}$	1.0	10.7	93.7	4.5	10.1	89.1	9.5	16.8	91.6	1.7	9.3	93.0	3.7	12.8	93.7
$\bar{y}_{k4.s}$	0.9	10.5	93.8	4.6	10.1	89.0	9.0	16.0	91.8	1.8	9.3	92.8	3.5	12.3	93.9
CONCAVE															
\bar{y}_r	0.9	23.7	90.9	19.0	31.6	92.3	15.0	26.5	76.1	33.2	41.7	76.4	37.1	41.4	37.5
$\bar{y}_{r\text{raim}}$	0.6	21.4	90.6	5.8	21.7	92.8	7.0	22.1	85.6	14.0	25.0	90.0	27.6	33.5	52.0
$\bar{y}_{c2.s}$	1.2	21.1	91.8	0.4	19.8	91.8	2.0	22.2	92.4	7.3	20.8	93.4	17.8	28.2	71.7
$\bar{y}_{c4.s}$	1.2	21.3	91.5	0.3	19.9	91.5	1.8	22.3	92.4	6.7	20.6	93.4	18.5	28.5	70.5
$\bar{y}_{k2.s}$	1.6	21.2	91.9	3.0	21.0	92.0	3.0	22.2	92.6	9.8	22.7	91.7	16.2	27.6	74.0
$\bar{y}_{k4.s}$	1.4	21.3	91.6	2.9	21.0	91.8	2.6	22.0	92.3	9.5	22.7	91.7	17.6	27.7	72.6
NON-QUOTIENT															
\bar{y}_r	0.1	10.7	92.9	9.7	14.6	86.5	7.3	12.6	81.3	11.9	16.1	80.8	8.8	13.5	77.8
$\bar{y}_{r\text{raim}}$	0.2	9.6	94.5	2.1	9.5	92.4	2.6	10.5	95.3	2.1	9.6	94.4	1.6	9.9	93.3
$\bar{y}_{c2.s}$	1.1	11.4	92.5	7.0	11.9	83.5	11.9	18.8	89.2	2.6	10.0	90.9	5.3	14.5	92.5
$\bar{y}_{c4.s}$	1.0	11.3	92.4	7.3	12.1	82.8	11.5	18.1	89.4	2.7	10.1	90.6	4.9	13.8	92.7
$\bar{y}_{k2.s}$	1.3	11.2	93.4	5.0	10.9	86.9	11.3	19.0	90.7	1.3	9.6	92.8	4.7	14.3	93.5
$\bar{y}_{k4.s}$	1.1	10.9	93.4	5.2	11.1	86.5	10.6	17.8	91.1	1.3	9.7	92.6	4.1	13.4	93.8

5.1 Régression de type QUOTIENT

L'étude du tableau 2 nous permet de tirer une série de conclusions.

(i) Le mécanisme (M1) (non-réponse uniforme).

Quand le mécanisme (M1) s'applique, l'estimateur non corrigé $\bar{y}_{r\text{raim}}$ est essentiellement sans biais et aucune correction n'a à être apportée. Toutefois, si l'analyste, qui soupçonnait l'existence d'un mécanisme confondu, a néanmoins choisi un des estimateurs corrigés, la conséquence n'est pas grave. Les huit estimateurs corrigés ne montrent qu'une petite augmentation de la VABR et de la RCEQMR comparativement à $\bar{y}_{r\text{raim}}$.

(iii) Les mécanismes (M2) et (M3) (non-confondu, non uniformes et dépendants de la valeur x).

Pour ces mécanismes, on voit que la VABR est très faible pour l'estimateur non corrigé $\bar{y}_{r\text{raim}}$ ce que la théorie nous aurait laissé prévoir. Nous nous concentrons plutôt sur le comportement des huit estimateurs corrigés, puisqu'il est important de savoir si des conséquences négatives sont associées à la décision d'utiliser, à tort, un de ces estimateurs. Une telle décision découlerait de l'hypothèse inexacte

nisme (M2) que pour le mécanisme (M3).

(iii) Le mécanisme (M4) (confondu et dépendant des valeurs y).

Pour ce mécanisme, une caractéristique remarquable du tableau 2 est le fait que les huit estimateurs corrigés entraînent tous une réduction appréciable du biais comparativement à l'estimateur non corrigé $\bar{y}_{r\text{raim}}$ (et une réduction très appréciable par rapport à l'estimateur élémentaire \bar{y}_r). On observe aussi, pour les estimateurs corrigés, une certaine amélioration dans la RCEQMR comparativement à $\bar{y}_{r\text{raim}}$. Les estimateurs corrigés en s donnent de meilleurs résultats que les estimateurs corrigés en r . Dans le groupe des estimateurs corrigés en s , les différences sont mineures, comme c'est le cas dans le groupe des estimateurs corrigés en r .

$$\text{RCEQMR}(\underline{y}) = 100 \times \frac{\sqrt{E_p E_q(\underline{y}^U - \underline{y}^U)^2}}{\underline{y}^U}.$$

Les espérances $E^p E^q(\hat{y}^U)$ et $E^p E^q(\hat{y}^U - \bar{y}^U)^2$ ont été

ensembles des répondants obtenus pour chacune des 240 combinaisons. Avec ce nombre de répétitions, l'erreur de Monte Carlo était inférieure à 0,1%, si l'on suppose que la distribution des \hat{p} est approximativement normale. Nous utiliserons l'abréviation "VABR" pour désigner la valeur absolue du biais relatif, $|BR(\hat{p})|$.

Nous traiterons aussi du niveau de confiance réel ("NCR"), de l'intervalle à 95% construit de la façon suivante:

(5.1)

$$(1-w)/z(\underline{d}-\chi\mathcal{A}) \sum \left(\frac{N}{1} - \frac{w}{1} \right) = (\underline{d})\mathcal{A}$$

Pour la suite de l'analyse, nous répartissons les estimateurs corrigés en deux groupes: les estimateurs corrigés en y_{it} , qui sont basés sur des facteurs de correction dans lesquels

(Pour chaque mécanisme, on a calculé la moyenne de 12 cas formés par la combinaison de quatre taux de non-réponse et de trois niveaux de corrélation)

	M1			M2			M3			M4			M5		
	(uniforme)			(décroissant en x)			(croissant en x)			(décroissant en y)			(croissant en y)		
	BR	MIR	NCR	BR	MIR	NCR	BR	MIR	NCR	BR	MIR	NCR	BR	MIR	NCR
μ_r	0.2	13.9	92.5	12.9	19.1	86.0	9.5	16.5	81.1	19.1	23.6	72.3	14.9	19.9	68.2
μ_{ramp}	0.2	12.3	92.7	0.6	11.8	93.0	0.4	12.9	92.4	5.3	13.0	92.5	6.0	13.9	85.6
$\mu_{c2.s}$	1.0	13.3	92.4	4.4	12.6	88.9	8.9	18.3	93.0	1.8	11.8	92.4	3.6	15.3	92.2
$\mu_{c4.s}$	0.9	13.2	92.3	4.7	12.6	88.6	8.4	17.7	93.0	1.7	11.7	92.3	3.4	14.9	92.2
$\mu_{c2.s}^*$	1.1	13.2	92.8	2.4	12.0	90.9	8.0	18.5	93.5	1.7	11.7	93.3	2.2	15.3	92.0
$\mu_{c4.s}^*$	1.0	13.1	92.7	2.6	12.0	90.8	7.3	17.7	93.5	1.6	11.7	93.2	1.8	14.7	91.9
$\mu_{c1.s}$	1.7	14.7	91.4	5.9	13.4	86.4	15.7	26.2	87.6	1.9	12.2	90.9	8.9	21.3	89.8
$\mu_{c3.s}$	1.6	14.4	91.4	6.2	13.5	86.1	14.9	25.1	87.8	2.1	12.2	90.7	8.3	20.4	90.0
$\mu_{c1.s}^*$	2.0	14.7	92.3	3.1	12.3	90.0	15.9	29.6	88.9	1.1	11.7	92.8	8.3	23.8	90.7
$\mu_{c3.s}^*$	1.7	14.3	92.3	3.2	12.4	89.8	14.6	27.6	89.3	1.0	11.7	92.7	7.1	21.9	91.0

où

$$A_o = \frac{1}{n-1} \left\{ \sum_o X_k - \frac{\sum_o X_k^2}{\sum_o X_k} - \frac{\sum_o X_k}{\sum_o X_k} + \frac{\sum_o X_k}{\sum_o X_k} \right\},$$

$$A_1 = \frac{X_s X_o}{X_r}$$

et

$$\hat{\sigma}^2 = \frac{\sum_r e_k^2 / (m-1)}{r \{1 - (cv_{X_r})^2 / m\}}, \quad (3.2)$$

où

$$e_k = y_k - B_{rX_k} - B_{rX_k, cv_{X_r}} = \frac{\sum_r (X_k - \bar{X}_r) / (m-1)}{X_r}.$$

La variance de \hat{y}^{raimp} a deux composantes, nommément,

la variance d'échantillonnage et la variance attribuable à l'imputation. Le premier terme de (3.1) (désigné par V^{ord}) est une estimation de la variance d'échantillonnage calculée à l'aide de la formule de la variance ordinaire en supposant que les données imputées sont aussi bonnes que les observations réelles. Puisque cette hypothèse n'est pas vérifiée, V^{ord} sous-estime la vraie variance d'échantillonnage. Pour corriger cette sous-estimation, on ajoute le deuxième terme, V^{dif} , de (3.1). Le dernier terme de (3.1), V^{imp} , est une estimation de la variance attribuable à l'imputation.

Si nous calculons la moyenne des valeurs y à partir de l'ensemble de données complet $\{y^c_k; k \in s\}$ défini en (2.6), nous obtenons l'estimateur (2.7). L'estimateur de la variance de ce dernier devrait tenir compte du facteur de correction C . Si nous pouvons supposer que l'espérance $E_{\xi} E_p E_q$ est égale à $E_p E_q E_{\xi}$ (cela est vrai quand il y a non-réponse non-confondue), nous pouvons utiliser la méthode proposée par Särndal (1990, 1992) pour obtenir un estimateur de la variance qui tient compte de C . Toutefois, nous sommes surtout intéressés par les cas où le mécanisme de réponse est confondu. Nous proposons donc un estimateur de la variance basé sur l'argument heuristique suivant.

L'estimateur $\hat{\sigma}^2$ dans (3.2) n'utilise que les données des répondants. Il sera certainement biaisé pour des mécanismes confondus et il faut apporter une correction si l'on veut utiliser la formule (3.1) pour l'estimateur corrigé (2.7). Nous posons de remplacer $\hat{\sigma}^2$ dans (3.1) par $C^2 \hat{\sigma}^2$, afin d'obtenir l'estimateur de la variance ci-après pour l'estimateur $\hat{y}^{c,s}$ dans (2.7):

$$V(\hat{y}^{c,s}) = V^{\text{ord}} + C^2(V^{\text{dif}} + V^{\text{imp}}), \quad (3.3)$$

où V^{ord} est calculée à l'aide des données après imputation avec le facteur de correction du biais C . Si l'on remplace C^2 par c_k^2 ou par k_k^2 , nous obtenons les estimateurs de la variance correspondant à $\hat{y}^{c,s}$ ou à $\hat{y}^{kt,s}$. Ces estimateurs donnent de très bons résultats dans un grand nombre des cas sur lesquels a porté la simulation mentionnée dans la section 5.

4. ETUDE DE SIMULATION

Nous considérons huit estimateurs corrigés correspondant aux huit facteurs de correction donnés dans (2.8) et (2.11). Une étude de simulation a été réalisée afin de déterminer si les estimateurs corrigés parvenaient bien à rétablir \hat{y}_s suivant différents mécanismes de réponse, en particulier, des mécanismes confondus. À des fins de comparaison, nous avons aussi inclus les estimateurs non corrigés \hat{y}_r et $\hat{y}^{\text{raimp}} = \bar{X}_s \bar{y}_r / \bar{X}_r$ donnés en (2.2). Notre objectif principal était d'examiner les estimateurs corrigés quand la population finie suit le modèle d'imputation par quotient défini en (2.3). Toutefois, nous désirions aussi voir comment les estimateurs corrigés se comportent dans des relations autres que la régression linéaire passant par l'origine. Nous avons aussi étudié les niveaux de confiance réels associés aux différents estimateurs quand les intervalles de confiance sont calculés à l'aide des estimateurs de la variance proposés dans la section 3.

Pour la simulation, nous avons créé 12 populations finies différentes, chacune de taille $N = 100$, en définissant de différentes façons les constantes a , b , c , et d dans le modèle de régression:

$$\mathbb{E}: y_k = a + bx_k + cx_k^2 + \epsilon_k, \quad E_{\mathbb{E}}(\epsilon_k) = 0,$$

$$V_{\mathbb{E}}(\epsilon_k) = d^2 x_k, \quad (4.1)$$

où l'on suppose que les ϵ_k sont indépendants. Quatre types de régression différents ont été créés à l'aide de quatre spécifications différentes de (a, b, c) . Ces types de régression sont appelés QUOTIENT ($a = c = 0, b > 0$), en conformité avec le modèle d'imputation par quotient défini dans (2.3), CONCAVE ($a = 0, b > 0, c < 0$), CONVEXE ($a = 0, b > 0, c > 0$) et NON-QUOTIENT ($a \neq 0, b > 0, c = 0$). Pour chaque type de régression, on a déterminé trois valeurs pour le coefficient de corrélation de modèle ρ_{xy} , soit 0.7, 0.8 et 0.9, en choisissant une valeur appropriée pour d . On a obtenu ainsi 12 spécifications de (a, b, c, d) , comme on le voit dans le tableau 1. Pour chacune des 12 spécifications, nous avons produit 100 valeurs pour la population, (y_k, x_k) , $k = 1, \dots, 100$ à l'aide d'une méthode à deux étapes. Nous avons employé la distribution Γ , avec paramètres α et β la fonction de densité de cette distribution est

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) \quad \text{pour } x > 0. \quad (4.2)$$

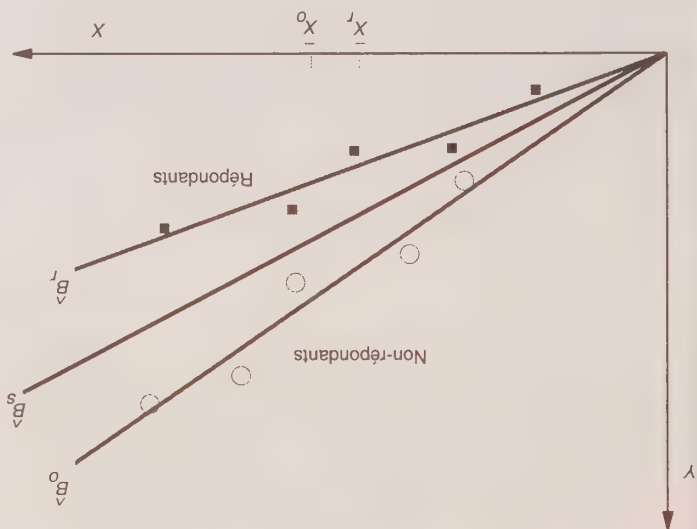


Figure 1. Exemple de représentation graphique des données (y_k, x_k) pour un mécanisme de réponse confondu.

Si l'on suppose que $C^{\text{opt}} > 1$, une méthode que peut appliquer l'analyste travaillant selon l'hypothèse III consiste à choisir une valeur C calculable qui satisfera probablement à la condition $C > 1$, puis à utiliser cette valeur C pour construire l'estimateur (2.7). Des facteurs C qu'on peut parfois utiliser de cette façon sont:

$$c_1 = \frac{x_r}{x_o}, \quad c_2 = \frac{x_s}{x_o}, \quad c_3 = \frac{w_r}{w_o}, \quad c_4 = \frac{w_s}{w_o}. \quad (2.8)$$

Ils sont basés sur le raisonnement que si le mécanisme de réponse est confondu tel que la probabilité de non-réponse est une fonction de y (par exemple, $\Theta_k = 1 - e^{-\gamma y_k}$ avec $\gamma > 0$), alors tant $C^{\text{opt}} > 1$ que $x_o > x_r$ sont vraisemblables, comme le montre la figure 1. Inversement, si la non-réponse est une fonction décroissante de y_k alors tant $C^{\text{opt}} < 1$ que $x_o < x_r$ sont vraisemblables. Une caractéristique importante de ces facteurs de correction est qu'ils peuvent être calculés au cours de la phase d'imputation mais qu'ils n'ont pas à l'être nécessairement. Par exemple, si l'on effectuait l'imputation par quotient habituelle $B_r x_k$ au cours de la phase d'imputation, il serait alors possible de calculer un facteur de correction approprié pendant la phase d'estimation sans changer les valeurs préalablement imputées. Remarque que c_2 suppose une correction un peu moins importante que c_1 : si $c_1 > 1$ nous avons $1 < c_2 < c_1$. Les choix $C = c_3$ et $C = c_4$ sont calculés sur les rangs des valeurs x , plutôt que sur les valeurs x proprement dites, afin d'amortir l'effet des valeurs x extrêmes. Plus précisément, soit w_k le rang de x_k dans l'ensemble de données $\{x_k: k \in s\}$. Les moyennes w en c_3 et en c_4 sont $w_s = (1/n) \sum_s w_k$, $w_r = (1/m) \sum_r w_k$ et $w_o = (1/l) \sum_o w_k$. Les quatre estimateurs obtenus en posant $C = c_i$ dans (2.7)

3. ESTIMATION DE LA VARIANCE

Puisque nous nous intéressons aux estimateurs de la variance basés sur une imputation simple, la méthode d'estimation de la variance proposée dans Särndal (1990, 1992) est intéressante. Si l'on suppose une non-réponse non-confondu et que le modèle ξ dans (2.3) est vérifié, l'estimateur de la variance de l'estimateur ponctuel \bar{y}^{raimp} dans (2.4) obtenu par cette méthode est donné par:

$$V(\bar{y}^{\text{raimp}}) = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{\sum_s (y_{\cdot k} - \bar{y}_{\cdot s})^2}{n - 1} + \left(\frac{1}{n} - \frac{1}{N} \right) A_o \hat{\sigma}_2^2 + \left(\frac{1}{n} - \frac{1}{N} \right) A_1 \hat{\sigma}_2^2 = V^{\text{ord}} + V^{\text{diff}} + V^{\text{imp}} \quad (3.1)$$

Les facteurs de correction présentés dans (2.8) ne donnent pas un résultat idéal quand le coefficient de corrélation entre x et y est proche de 1. Dans ce cas, nous avons $B_r \approx B_s \approx B_o$ pourvu que le modèle (2.2) soit vérifié. Par conséquent, le facteur de correction C devrait être proche de 1. Toutefois, les facteurs de correction présentés dans (2.8) pourraient être très différents de 1 et leur utilisation introduirait d'un biais. C'est pourquoi il peut être préférable d'utiliser dans (2.7) un facteur de correction C qui tient compte de la corrélation. Des facteurs de correction de ce genre sont:

$$k_i = 1 - \{ (c_i^2 - 1) (R_{xy}^2 - 1) \}, \quad i = 1, \dots, 4. \quad (2.11)$$

où c_i , $i = 1, \dots, 4$ désigne les quatre facteurs de correction présentés dans (2.8) et R_{xy} est le coefficient de corrélation estimé fondé sur les données des répondants. Dans notre simulation de Monte Carlo, nous avons aussi inclus l'estimateur (2.7) correspondant aux quatre choix $C = k_i$, $i = 1, \dots, 4$. Ces estimateurs seront désignés par $\bar{y}^{k_i, s}$, $i = 1, \dots, 4$.

$$\bar{y}^{c_2, s} = \bar{y}_r \left[1 + \left(1 - \frac{n}{m} \right) \left\{ \frac{x_o^2}{x_r x_s} - 1 \right\} \right]. \quad (2.10)$$

$$\bar{y}^{c_1, s} = \bar{y}_r \left[1 + \left(1 - \frac{n}{m} \right) \left\{ \left(\frac{x_r}{x_o} \right)^2 - 1 \right\} \right], \quad (2.9)$$

selon (2.8) seront désignés par $\bar{y}^{c_i, s}$, $i = 1, \dots, 4$. En particulier, nous avons

Hypothèse III: (III-1): le mécanisme de réponse est arbi-

traire sauf pour le fait qu'il est

confondu;

(III-2): le modèle de l'imputation par

quotient (2.2) est vérifié.

Il est habituellement difficile sinon impossible pour

l'analyste de déterminer laquelle des hypothèses II ou III est la plus appropriée. L'examen des données n'aidera pas beaucoup si les seules données disponibles se rapportent

au moment présent, ce qui est généralement le cas dans une enquête unique. Il est alors impossible de vérifier l'hypo-

thèse faite (qu'il s'agisse de l'hypothèse II ou de l'hypo-

thèse III). Par contre, si l'analyste possède de l'expérience

relativement à une enquête répétée régulièrement, cette

personne peut alors avoir des raisons légitimes de croire,

par exemple, que la non-réponse est une fonction de la

variable étudiée.

Dans certaines situations, on peut faire l'hypothèse

qu'il existe un mécanisme confondu pour les raisons

suitantes. Supposons que, dans une enquête sur les finances

personnelles, la variable à l'étude y est l'"épargne" et que

la variable auxiliaire x est le "revenu", les valeurs x_k étant

connues pour les individus $k \in s$. Il est vraisemblable qu'il

existe une corrélation entre la probabilité de non-réponse

du répondant k et le montant de l'épargne y_k qu'on lui

demande de divulguer ainsi que le montant du revenu x_k

que l'on connaît d'autres sources. Mais puisque l'épargne,

et non le revenu, est la variable que le répondant doit

divulguer directement dans le cadre de l'enquête, il peut

être plus réaliste de supposer que la probabilité de non-

réponse est une fonction de y_k que de supposer qu'elle est

une fonction de x_k . Par conséquent, il pourrait être plus

réaliste de supposer qu'il existe un mécanisme confondu

plutôt qu'un mécanisme non-confondu.

Selon l'hypothèse III, ni y_r ni y_r^{ramp} ne sont des estima-

teurs de rétablissement. On peut exprimer le Biais-C de

$$\text{Biais-C}(y^{\text{ramp}}) = x_s E_{\xi} E_{\eta} \left(\frac{\sum_r^r x_k}{\sum_r^r e_k} \right),$$

y^{ramp} par la formule

Biais-C(y^{ramp}) = $x_s E_{\xi} E_{\eta} \left(\frac{\sum_r^r x_k}{\sum_r^r e_k} \right)$.

On peut exprimer le Biais-C de

pour les répondants $\sum_r^r e_k$ tend à être négatif.

Un mécanisme de réponse confondu (comme dans

l'hypothèse III) introduit un biais dans l'estimateur de la

pente $B_r = (\sum_r^r y_k) / (\sum_r^r x_k)$. Par conséquent, B_{r,x_k} est

une imputation biaisée pour une valeur manquante y_k .

Afin d'améliorer la situation, supposons qu'une valeur

manquante y_k est imputée par $C B_{r,x_k}$ plutôt que par B_{r,x_k} où C est une quantité à préciser. Alors, les données après

imputation sont définies par

$$y_k^c = \begin{cases} y_k, & \text{si } k \in r \\ C B_{r,x_k}, & \text{si } k \in o \end{cases} \quad (2.6)$$

et, désignant la moyenne d'échantillon de ces données par

$$y_{c,s} = (1/n) \sum_s y_k^c, \text{ nous obtenons l'estimateur}$$

$$y_{c,s} = y_r \left[1 + \left(1 - \frac{n}{m} \right) \left(C \frac{x_r}{x_o} - 1 \right) \right]. \quad (2.7)$$

Une correction simple comme celle utilisée dans (2.6) a été

mentionnée dans Rubin (1986, 1987, p. 203) dans le con-

texte de l'imputation multiple. Rubin voit C comme une

constante choisie par l'utilisateur selon les connaissances

de la constante est juste, le biais de (2.7) peut être faible.

Ici, nous examinerons des choix de C qui sont adap-

ratifs, c'est-à-dire qui reflètent l'échantillon obtenu s et

l'ensemble de répondants réalisé r . Idéalement, C devrait

être tel que l'imputation rétablira exactement l'estimateur

$y_s = (1/n) \sum_s y_k$ qui serait utilisé avec un taux de

réponse de 100%. Cette valeur de C est déterminée par

l'équation

$$y_s = \frac{1}{n} \sum_s y_k = \frac{1}{n} \sum_s y_k^c = \frac{1}{n} \left(\sum_r^r y_k + \sum_o^o C B_{r,x_k} \right).$$

Un calcul simple montre que la valeur optimale de C est

$$C^{\text{opt}} = \frac{B_r}{B_o},$$

où $B_o = \sum_o y_k / \sum_o x_k$ est l'estimation de la pente si le

modèle (2.2) pouvait être ajusté aux non-répondants. Les

valeurs imputées seraient alors $y_k = B_o x_k$ pour $k \in o$.

Manifestement, C^{opt} et B_o ne peuvent être calculés puis-

qu'ils dépendent de valeurs manquantes de y_k . Pour un

mécanisme non-confondu (comme dans l'hypothèse II),

nous pouvons nous attendre que $C^{\text{opt}} \approx 1$ étant donné s

parce que

$$E_{\xi} E_{\eta} (C^{\text{opt}} | s) = E_{\eta} E_{\xi} \left(\frac{B_r}{B_o} | s \right) \approx 1.$$

Mais pour un mécanisme confondu (comme dans l'hypo-

thèse III), C^{opt} peut être très loin de l'unité. Supposons

que $C^{\text{opt}} > 1$. Remarquez que $C^{\text{opt}} > 1$ si et seulement si

$\sum_r^r e_{ks} > 0$, avec $e_{ks} = y_k - B_{s,x_k}$, où $B_s = (\sum_s y_k) /$

$(\sum_s x_k)$ est l'estimation inconnue de la pente pour un taux

de réponse de 100%. C'est-à-dire que $C^{\text{opt}} > 1$ implique

que les résidus e_{ks} pour les répondants sont négatifs en

moenne. Une illustration de cette situation est présentée

dans la figure 1, où $n = 10$, $l = n - m = 5$ et les résidus

e_{ks} des cinq répondants sont tous négatifs.

Voici un exemple de mécanisme de réponse non-confondu

$$q(r | s) = \prod_{k \in r} (1 - \theta_k) \prod_{k \in s-r} \theta_k,$$

où $\theta_k = 1 - P(k \in r | s) = 1 - e^{-\gamma_{rk}}$, pour une constante positive γ , est la probabilité de non-réponse de l'unité k . Par contre, si $\theta_k = 1 - e^{-\gamma_{rk}}$, alors $q(r | s)$ est un mécanisme confondu.

Le mécanisme de réponse uniforme défini par $q(r | s) = (1 - \theta)^{m\Theta_n - m}$ est un mécanisme non-confondu particulièrement simple. Ici, les unités répondent selon des expériences de Bernoulli $(1 - \theta)$ indépendantes et identiques, où θ est la probabilité de non-réponse commune à toutes les unités.

Le fait de considérer qu'un estimateur d'imputation \hat{y}_U de y_U , y compris y^{ramp} défini en (2.4), est bon dépend en partie des hypothèses faites par l'analyste à propos du mécanisme de réponse et en partie de la relation entre y et x . Plus loin dans cette section nous traitons de plusieurs hypothèses possibles. Pour tout s donné, l'objectif est que, suivant des hypothèses réalistes précises, l'espérance de la différence $\hat{y}_U - y_s$ soit proche de zéro. Autrement dit, suivant les hypothèses données, le biais conditionnel (Biais-C) de \hat{y}_U , Biais-C(\hat{y}_U) = $E(\hat{y}_U - y_s | s)$ devrait être faible. Nous appelons \hat{y}_U un *estimateur de réajustement* de y_U si Biais-C(\hat{y}_U) = 0 ou ≈ 0 c'est-à-dire, si l'espérance conditionnelle de \hat{y}_U est (approximativement) égale à y_s . Il s'ensuit que si le Biais-C est (approximativement) égal à zéro pour tout s , alors le biais inconditionnel, soit sur tous les échantillons possibles est aussi (approximativement) égal à zéro.

Les hypothèses varient selon les analystes. Étudions certaines hypothèses courantes et posons-nous la question: Quels estimateurs de réajustement ces hypothèses permettent-elles d'utiliser?

Hypothèse I: Le mécanisme de réponse est uniforme.

En vertu de l'hypothèse I, y^{ramp} est un estimateur de réajustement. Pour constater que tel est bien le cas, remarquons que

$$\text{Biais-C}(y^{\text{ramp}}) = E_q(y^{\text{ramp}} | s) - y_s \approx 0,$$

parce que, étant donné s , y^{ramp} est l'estimateur par quotient classique de y_s . L'hypothèse I n'est pas réaliste dans la plupart des enquêtes. On sait que la propension à répondre varie en fonction de caractéristiques observables telles que la taille et la branche d'activité (pour les établissements commerciaux), la taille de la famille et le genre de famille (pour les ménages), l'âge, le sexe et le revenu (pour les particuliers). Selon cette hypothèse irréaliste, même un estimateur élémentaire comme la moyenne des répondants $y_r = (1/m) \sum_r y_k$ est un estimateur de réajustement:

$$\text{Biais-C}(y_r) = E_q(y_r | s) - y_s = 0.$$

Toutefois, si l'hypothèse I est vérifiée, y^{ramp} est préféré à y_r parce que l'estimateur par quotient est caractérisé par une variance plus faible si le modèle ξ est valide.

L'analyste doit manifestement considérer des hypothèses plus réalistes qui permettent aux probabilités de réponse de varier selon les caractéristiques observables. L'hypothèse ci-après, composée de deux parties, est de ce genre.

Hypothèse II: (II-1): le mécanisme de réponse est arbitraire sauf pour le fait qu'il est non-confondu;

(II-2): le modèle de l'imputation par quotient (2.2) est vérifié.

Ici, (II-1) est une exigence moins stricte et plus réaliste relativement au mécanisme de réponse que l'exigence d'uniformité de l'hypothèse I. Selon (II-1), le mécanisme de réponse peut avoir n'importe quelle forme pourvu qu'il soit non-confondu. Toutefois, les hypothèses I et II ne sont pas directement comparables puisque l'hypothèse II contient une composante de modèle, (II-2), qui manque dans l'hypothèse I. Selon l'hypothèse II, y^{ramp} est un estimateur de réajustement parce que

$$\begin{aligned} \text{Biais-C}(y^{\text{ramp}}) &= E_{\xi}\{E_q(y^{\text{ramp}}) - y_s | s\} \\ &= E_{\xi} E_q \left(E_{\xi} \left(\frac{y_r}{x_s} \right) - E_{\xi}(y_s) \right) \\ &= E_q(\beta x_s) - \beta x_s = 0. \end{aligned}$$

Remarquez que l'on peut changer l'ordre des espérances, $E_{\xi} E_q$ à $E_q E_{\xi}$ suivant l'hypothèse II, parce que le mécanisme de réponse est alors de la forme $q(r | x_r)$ c'est-à-dire qu'il ne dépend pas des valeurs y . Par contre, la moyenne des répondants y_r n'est pas un estimateur de réajustement parce que

$$\text{Biais-C}(y_r) = E_{\xi}\{E_q(y_r) - y_s | s\} = \beta\{E_q(x_r) - x_s\},$$

qui est généralement non nul selon l'hypothèse II. Nous pouvons, toutefois, transformer y_r en un estimateur de réajustement à l'aide d'un facteur de correction multiplicatif. Cela donne

$$y_r \left\{ 1 + \left(1 - \frac{m}{n} \right) \left(\frac{x_r}{x_o} - 1 \right) \right\}, \quad (2.5)$$

qui n'est qu'une autre façon d'exprimer y^{ramp} comme on peut facilement le vérifier. Dans un exemple qui utilise l'approche bayésienne, Little et Rubin (1987, p. 233) obtiennent un estimateur identique à l'estimateur (2.5). Étudions maintenant les mécanismes de réponse confondus. Ces derniers causent des problèmes plus difficiles quand on veut trouver un estimateur de réajustement.

Dans le cas des données de type nominal, on a aussi proposé quelques méthodes pour traiter le problème de la non-réponse non-ignorable. Par exemple, Baker et Laird (1988) essaient de modéliser le mécanisme de réponse à l'aide de modèles log-linéaires. On traite aussi de la modélisation causale dans Fay (1986, 1989).

À Statistique Canada, on utilise souvent l'imputation par quotient, particulièrement dans les enquêtes à passages répétés. Par exemple, dans l'Enquête mensuelle sur les industries manufacturières, on impute des valeurs pour les livraisons de la période courante à l'aide de la méthode d'imputation par quotient en utilisant les livraisons du mois précédent comme variable auxiliaire. Cette méthode simple est très intéressante pour les spécialistes parce qu'elle reflète la variation d'un mois à l'autre.

Dans le présent article, nous étudions la possibilité d'améliorer à l'aide de facteurs de correction simples l'estimateur appliqué à des données comprenant des valeurs imputées par quotient. Par conséquent, nous supposons que l'imputation a déjà été effectuée et nous essayons de corriger l'estimateur. Nous nous concentrons sur l'estimation de la moyenne. L'utilisation de facteurs de correction simples serait très intéressante pour l'utilisateur pourvu qu'elle donne d'assez bons résultats. Une telle procédure est aussi facile à appliquer sans avoir recours à un travail informatique considérable et elle nous évite de modéliser explicitement le mécanisme de non-réponse. Toutefois, nous utilisons des facteurs de correction dépendants de l'échantillon plutôt qu'une constante choisie *a priori*.

Dans la section 2, nous définissons plusieurs facteurs de correction simples qui répondent à nos exigences. Dans la section 3, nous proposons un estimateur de la variance qui peut être utilisé conjointement avec les estimateurs ponctuels corrigés. Nous avons examiné les propriétés des estimateurs ponctuels corrigés à l'aide d'une simulation de Monte Carlo dont nous traitons dans les sections 4 et 5. Dans la section 6, nous présentons des conclusions.

2. FACTEURS DE CORRECTION DU BIAIS SIMPLES

Désignons par $U = \{1, \dots, k, \dots, N\}$ l'ensemble des indices d'une population finie et représentons la moyenne de population de la variable étudiée y par $\bar{y}_U = (1/N) \sum_{U \in U} y_k$. Nous supposons que $y_k > 0$ pour tout $k \in U$. Nous tirons de U , sans remise, un échantillon aléatoire simple s de taille n (EASSR). L'estimateur non biaisé qui serait utilisé avec un taux de réponse de 100% est la moyenne de l'échantillon

$$\bar{y}_s = (1/n) \sum_{k \in s} y_k. \quad (2.1)$$

Soit r et o les ensembles des unités répondantes et non répondantes, respectivement, de sorte que $s = r \cup o$. Nous désignons le plan d'échantillonnage EASSR par $p(\cdot)$ et le mécanisme de réponse, étant donné s , par $q(\cdot | s)$.

Autrement dit, $p(s)$ est la probabilité que l'EASSR soit tiré et $q(r | s)$ est la probabilité que l'ensemble des répondants r soit réalisé étant donné l'échantillon s . Désignons aussi par m et par l la taille de r et de o , respectivement. Pour simplifier, nous supposons que la probabilité que $n = 0$ est négligeable. Nous supposons que l'imputation est effectuée à l'aide d'une variable auxiliaire, x , dont la valeur, x_k , est connue et positive pour tout $k \in s$. Si $k \in o$, la valeur manquante y_k est imputée par \hat{y}_k . L'ensemble de données complet est désigné par $\{y_k : k \in s\}$, où $y_k = y_k$ si $k \in r$ et $y_k = \hat{y}_k$ si $k \in o$.

Dans le présent article, nous étudions l'imputation par quotient. Cette méthode d'imputation courante est basée sur un modèle simple. C'est-à-dire que, si la valeur y_k manque, on l'impute par \hat{y}_k , où $\hat{y}_k = (\sum_{r \in r} y_r) / (\sum_{r \in r} x_r)$. Le modèle désigné par \hat{y}_k précise que, pour $k \in s$,

$$y_k = \beta x_k + \epsilon_k, \quad E_\xi(\epsilon_k | x_k) = 0, \quad V_\xi(\epsilon_k | x_k) = \sigma^2 x_k, \quad E_\xi(\epsilon_k | x_k, x_l) = 0, \quad k \neq l. \quad (2.2)$$

Selon ce modèle, \hat{y}_k est le meilleur prédicteur linéaire non biaisé de la valeur manquante y_k , d'après les données fournies par les répondants $\{y_k, x_k\} : k \in r\}$. L'ensemble de données complet est alors composé des valeurs

$$y_k = \begin{cases} y_k, & \text{si } k \in r \\ \hat{y}_k, & \text{si } k \in o. \end{cases} \quad (2.3)$$

La procédure habituelle consiste à prendre la formule de l'estimateur utilisée dans le cas d'un taux de réponse de 100% à l'ensemble de données complet. Cela donne

$$\bar{y}_s = \frac{1}{n} \sum_{k \in s} y_k = \frac{\bar{y}_r}{\bar{x}_r} \bar{x}_s = \bar{y}_{\text{ramp}}, \quad (2.4)$$

où $\bar{x}_s = (1/n) \sum_{k \in s} x_k$, $\bar{y}_r = (1/m) \sum_{r \in r} y_r$ et $\bar{x}_r = (1/m) \sum_{r \in r} x_r$. Remarquez que ramp signifie imputé par quotient. Nous devons maintenant chercher à savoir si l'imputation peut rétablir l'estimateur utilisé dans le cas d'un taux de réponse de 100%, \bar{y}_s , en ce sens que l'espérance de l'estimateur d'imputation \bar{y}_s est égale à \bar{y}_s , étant donné s . À moins que cela ne puisse se réaliser, l'imputation par quotient aura introduit un biais. Pour examiner cette question, nous devons étudier le mécanisme de réponse. Aux fins du présent article, nous disons qu'un mécanisme de réponse $q(\cdot | s)$ est *non-confondu* s'il est de la forme $q(r | s) = q(r | x_s)$ où $x_s = \{x_k : k \in s\}$ et les probabilités de réponse satisfont $P(k \in r | s) > 0$ pour tout $k \in s$. C'est-à-dire qu'il peut dépendre de s et des valeurs x qui lui sont associées. S'il dépend aussi des valeurs y , de sorte que $q(r | s) = q(r | x_s, y_s)$, alors on dit qu'il s'agit d'un mécanisme de réponse *confondu*. Dans ces définitions, le mécanisme de réponse dépend de l'échantillon obtenu s . On trouve des définitions légèrement différentes de 'mécanisme de réponse confondu' et de 'mécanisme de réponse non-confondu' dans Rubin (1987, p. 39), dans ce dernier cas, les mécanismes de réponse sont non-conditionnels.

Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme confondu

E. RANCOURT, H. LEE et C.-E. SÄRNDAI¹

RÉSUMÉ

La plupart des enquêtes souffrent du problème de données manquantes attribuable à la non-réponse. Pour traiter ce problème, on a souvent recouru à l'imputation afin de créer un "ensemble de données complet", c'est-à-dire, un ensemble de données composé d'observations réelles (pour les répondants) et d'imputations (pour les non-répondants). Habituellement, on effectue l'imputation en supposant un mécanisme de réponse non-confondu. Quand cette hypothèse se révèle fausse, un biais est introduit dans l'estimateur ordinaire de la moyenne de population calculé à partir de l'ensemble de données complet. Dans le présent article, nous étudions l'idée d'employer des facteurs de correction simples pour régler le problème du biais dans le cas où l'on a recouru à l'imputation par quotient. Nous évaluons l'efficacité des facteurs de correction à l'aide d'une simulation de Monte Carlo dans laquelle nous utilisons des ensembles de données produits artificiellement qui représentent divers taux de non-réponse et mécanismes de non-réponse et diverses superpopulations et corrélations entre la variable étudiée et la variable auxiliaire. Nous constatons que ces facteurs de correction sont efficaces, particulièrement lorsque la population suit le modèle sous-jacent à l'imputation par quotient. Nous traitons aussi d'une option pour estimer la variance des estimations ponctuelles corrigées.

MOTS CLÉS: Biais conditionnel; simulation Monte Carlo; estimateur de rétablissement; estimation de la variance.

1. INTRODUCTION

Dans les enquêtes, l'existence de la non-réponse est plutôt la règle que l'exception. On impute souvent les données manquantes attribuables à la non-réponse afin d'obtenir un ensemble de données complet et on utilise l'estimateur ordinaire à partir de cet ensemble de données en supposant que le mécanisme de réponse sous-jacent est non-confondu. Toutefois, une estimation ponctuelle obtenue de cette façon est biaisée lorsque le mécanisme de réponse est confondu. Dans ce cas, le biais pourrait être très important, comme on le fait remarquer dans Lee, Rancourt et Särndal (1994). Selon Rubin (1987, p. 39), un mécanisme de réponse est non-confondu s'il ne dépend pas de la variable étudiée; autrement, il s'agit d'un mécanisme confondu. (Une définition formelle adaptée aux besoins du présent article sera présentée à la section 2.) Dans un cadre bayésien, le mécanisme de réponse non-confondu est appelé "ignorable". Pour un biais causé par un mécanisme de réponse non-ignorable, Rubin (1977, 1987) et Little et Rubin (1987) ont étudié une méthode qui permet de corriger la moyenne des répondants à l'aide de variables auxiliaires. Dans cette méthode, on suppose une régression linéaire entre la variable étudiée y et un vecteur de coefficients x . On suppose aussi que le vecteur des coefficients de régression pour les non-répondants suit une distribution normale *a priori* avec une moyenne égale au vecteur des coefficients de régression pour les répondants.

Après avoir supposé un modèle logistique pour la probabilité de réponse, Greenless, Reece et Zieschang (1982) ont proposé une méthode pour traiter la non-réponse non-ignorable à l'aide de l'estimation par la méthode du maximum de vraisemblance. De plus, on suppose qu'il existe un modèle de régression linéaire pour la relation entre y et x , un vecteur de variables auxiliaires. Le modèle logistique de la probabilité de réponse inclut y et z , un vecteur de variables auxiliaires différentes. En supposant aussi que le terme d'erreur de la régression suit une distribution normale, les auteurs précités obtiennent des estimations du maximum de vraisemblance des paramètres inconnus du modèle de régression et du modèle logistique. Finalement, pour un non-répondant, on calcule une valeur imputée comme la moyenne de la distribution conditionnelle de y étant donné les valeurs de x et z pour les non-répondants, et les paramètres estimés. Il se peut qu'une telle méthode donne de bons résultats quand toutes les hypothèses du modèle sont respectées mais il est probable qu'elle soit très sensible aux spécifications des deux modèles. On ne peut habituellement pas faire de test pour vérifier l'adéquation du modèle de la probabilité de réponse. Toutefois, si des données peuvent être obtenues d'une source externe, il peut alors être possible de faire un test portant sur le modèle de la probabilité de réponse, comme Greenless et coll. l'ont fait à l'aide de données provenant de la Current Population Survey. L'application de cette méthode exige des calculs intensifs.

¹ E. Rancourt et H. Lee, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6; C.-E. Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec), Canada, H3C 3J7.

REMERCIEMENTS

Cet article fait état de travaux de recherche effectués à contrat pour Statistique Canada. L'auteur souhaite signaler les travaux des consultants et entrepreneurs suivants: D.R. Harley Consultants Limited, Kennedy et de Groh Consultants et Price Waterhouse Conseillers en gestion. Les opinions exprimées ici sont celles de l'auteur et ne représentent pas nécessairement celles de Statistique Canada ou de ces entrepreneurs.

BIBLIOGRAPHIE

BUREAU, M. (1991). Experience with the use of cognitive methods in designing business survey questionnaires. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-717.

DEMAIO, T.J. (ED.) (1983). *Approaches to Developing Questionnaires*. Statistical Policy Working Paper 10, Washington, DC: United States Office of Management and Budget.

EDWARDS, W.S., et CANTOR, D. (1991). Toward a response model in establishment surveys. Dans *Measurement Errors in Surveys*. (Eds. Paul P. Biemer et coll.). New York: John Wiley and Sons, 211-233.

GROSS, GILROY et ASSOCIATES LTD. (1989). Qualitative Research to Evaluate the Questionnaire of the Survey of Employment, Payrolls and Hours (SEPH). Rapport final soumis à Statistique Canada.

GROSS, GILROY et ASSOCIATES LTD. (1990). Qualitative Research to Evaluate the Redesign Survey Materials of the Survey of Employment, Payrolls and Hours (SEPH). Rapport final soumis à Statistique Canada.

GOWER, A.R. (1993). Questionnaire design for establishment surveys. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 950-956.

GOWER, A.R. (1991). The Questionnaire Design Resource Centre's Role in Questionnaire Research and Development at Statistics Canada. 48ième session de l'Institut International de Statistique. *Recueil*, Volume III, 58-59.

GOWER, A.R., et NARGUNDKAR, M.S. (1991). Cognitive Aspects of Questionnaire Design: Business Surveys versus Household Surveys. *Proceedings of the 1991 Annual Research Conference*. Washington, DC: United States Bureau of the Census, 299-312.

GOWER, A.R., et ZYLSTRA, P.D. (1990). The Use of Qualitative Methods in the Design of a Business Survey Questionnaire. Contributed Paper (non-publié). International Conference on Measurement Errors in Surveys, Tucson, Arizona.

D.R. HARLEY CONSULTANTS LIMITED (1993). Qualitative Testing of the Draft National Training Survey Questionnaire. Rapport final soumis à Statistique Canada.

KENNEDY et DE GROH CONSULTANTS (1992). Testing of Definitions for the National Training Survey. Rapport final soumis à Statistique Canada.

NOONAN, M. (1992). Final report on Personal Interviews with Potential Respondents for the Proposed Wholesale and Retail Trades Survey. Rapport non-publié, Statistique Canada.

PRICE WATERHOUSE MANAGEMENT CONSULTANTS (1990). Qualitative Research Related to the Re-design of the Census of the Construction Industry Questionnaires. Rapport final soumis à Statistique Canada.

STATISTIQUE CANADA (1989). Construction Census Questionnaire Test. Rapport non-publié, Section du recensement de la construction, Division de l'industrie.

STATISTIQUE CANADA (1994). Politique sur l'élaboration, l'essai et l'évaluation des questionnaires.

STATISTIQUE CANADA (1986). Politique d'information des répondants aux enquêtes.

TOURANGEAU, R. (1984). Cognitive sciences and survey methods. Dans *Cognitive Aspects of Survey Methodology: Building a Gap Between Disciplines*. (Eds. T.B. Jabine, M.L. Straf, J.M. Tanur, et R. Tourangeau). Washington, DC: National Academy Press, 73-100.

Certains répondants ont proposé que les données soient groupées à l'échelle nationale et provinciale et par secteur afin qu'ils puissent se comparer à d'autres sociétés dans leur domaine et dans leur région du pays. Comme un répondant l'a déclaré: "Je voudrais que les données soient propres à notre branche d'activité, avec le volume et le genre de formation donnée... Nous pourrions nous comparer aux autres entreprises de notre secteur quant au nombre d'employés qui reçoivent une formation et au pourcentage de la masse salariale consacrée à la formation du personnel."

De nombreux répondants des petites et moyennes entreprises trouvaient le questionnaire trop général et le niveau de détail trop compliqué pour leur permettre de répondre. Selon eux, le questionnaire était conçu pour des organisations plus grandes. Par exemple, les répondants de nombreuses petites entreprises estimaient que les catégories du questionnaire ne convenaient pas à leur entreprise. Ils pensaient qu'une bonne partie de la formation qu'ils offraient à leur personnel appartenait à la catégorie "non structurée" et que le questionnaire ne recueillait pas de renseignements sur cet aspect de la formation. Mais d'autres répondants de petites et moyennes entreprises ont déclaré que le questionnaire était très complet.

Les grandes entreprises ont aussi éprouvé de la difficulté avec le niveau de détail requis pour l'enquête. Le principal problème était qu'elles classent leurs dossiers selon le type de formation plutôt que selon la catégorie professionnelle du personnel formé.

Dans l'ensemble, on a observé des pratiques diverses en matière de tenue des dossiers. Certaines entreprises tiennent d'excellents dossiers sur la formation, d'autres non. Les participants qui ne tenaient pas de bons dossiers ou dont les dossiers ne contenaient pas les renseignements voulus trouvaient qu'il était difficile de répondre au questionnaire. D'autres, qui avaient des dossiers très détaillés, pouvaient manipuler les données dont ils disposaient de manière à les adapter au questionnaire. La seule exception était les questions sur les dépenses de formation, pour lesquelles les entreprises avaient de la difficulté à donner des renseignements détaillés. Selon les répondants de ces entreprises, les chiffres globaux étaient plus faciles à obtenir. Beaucoup d'entreprises ont précisé que leurs dossiers de formation n'étaient pas centralisés, de sorte que pour elles le questionnaire était plus long et plus difficile à remplir. Les répondants de ces entreprises ont dit qu'ils rempliraient ce qu'ils pourraient puis coordonneraient le travail nécessaire pour remplir le reste du questionnaire en envoyant ce dernier dans plusieurs services de l'entreprise.

Même si au début de nombreux participants étaient découragés par la longueur et la complexité apparentes du questionnaire, ils l'ont trouvé plus facile à remplir qu'ils ne l'avaient prévu. De nombreux participants ont trouvé que le questionnaire était tellement complet qu'il les avait fait se rappeler beaucoup d'activités de formation sur lesquelles ils n'auraient pas normalement donné de renseignements.

La plupart des participants pensaient que le questionnaire devrait être plus court. Mais ils ont aussi proposé d'ajouter quelques autres questions à réponse libre sur la formation prévue. Pour ce qui est du fardeau de déclaration, les répondants (particulièrement dans les entreprises de moyenne et de grande taille) estimaient qu'il allait falloir des heures pour réunir l'information nécessaire pour répondre aux questions sur les dépenses de formation, sur les heures de formation et sur le nombre d'employés formés par catégorie professionnelle.

On a constaté des différences dans le temps dont les répondants avaient besoin pour remplir le questionnaire. Les petites entreprises prenaient entre dix minutes et une heure. Dans les grandes entreprises, on estimait qu'il fallait environ deux heures (D.R. Harley Consultants Limited).

6. CONCLUSIONS

Dans cet article, on a présenté un aperçu de la conception des questionnaires d'enquêtes-entreprises. Comme on l'a fait observer, de nombreuses considérations entrent en ligne de compte dans la conception de ces questionnaires. Ces considérations sont les objectifs et les besoins en données de l'enquête et la consultation des utilisateurs des données et des répondants sur les caractéristiques et les préoccupations des répondants. Les autres considérations sont le fardeau de déclaration, la méthode de collecte des données, la disponibilité des données, l'utilisation de dossiers et la nécessité de faire l'essai des questionnaires.

Les problèmes de conception dont il faut tenir compte sont les instructions, la précision et la lisibilité des questions, l'enchaînement logique des questions, la compatibilité des catégories de réponses et des périodes de référence avec les pratiques des répondants en matière de tenue des dossiers et les exigences du traitement des données. Le questionnaire doit être convivial pour les répondants comme pour les intervieweurs.

Dans les enquêtes-entreprises, il importe, pour assurer la collecte de données exactes et utiles, de comprendre le processus de réponse que suivent les enquêtés lorsqu'ils répondent à un questionnaire. Les groupes de discussion et les méthodes de recherche cognitives sont des moyens très efficaces pour étudier ce processus de réponse et pour faire l'essai des questionnaires. Ils donnent le moyen de consulter directement les répondants et, par conséquent, d'intégrer leurs idées, leurs préoccupations et leurs suggestions dans le processus de conception des questionnaires. Si nous nous tournons vers l'avenir, nous pouvons voir que la recherche et l'expérience devraient permettre d'améliorer les méthodes et les approches actuellement employées pour élaborer et mettre à l'essai les questionnaires d'enquêtes-entreprises. Un des domaines dans lesquels il faut pousser la recherche et le développement est le rapport entre le questionnaire et la source d'information externe et l'influence de ce rapport sur le processus de réponse et sur l'exactitude des données déclarées.

L'objet de l'ENF est de recueillir des renseignements sur la formation et le perfectionnement du personnel dans le secteur privé. On demande aux répondants de fournir des données sur le genre de formation et sur son volume, sur le nombre de stagiaires et sur les groupes professionnels auxquels ils appartiennent, sur les caractéristiques des entreprises qui assurent une formation à leurs employés et sur le montant consacré à cette activité. Dans les grandes entreprises, les répondants sont les personnes qui travaillent dans le service de planification et de formation des ressources humaines de leur société, alors que dans les entreprises plus petites c'est en général le propriétaire ou le chef de la direction.

Tôt dans le processus d'élaboration du questionnaire, des réunions de groupes de discussion ont eu lieu et des interviews en profondeur ont été réalisées avec des représentants de grandes, de moyennes et de petites sociétés. Ces méthodes ont été employées parce que Statistique Canada jugeait important de consulter des représentants des milieux d'affaires pour s'assurer que dans l'élaboration du questionnaire de l'ENF on tenait compte de leurs intérêts et de leurs préoccupations en matière de formation. Les groupes de discussion et les interviews ont permis d'évaluer la précision et la pertinence de la terminologie et des concepts associés à la formation des employés dans un établissement commercial. L'étude a porté sur le sens que donnaient les répondants à des termes tels que "formation régulière" (structurée) et "formation non structurée" et sur leur aptitude à utiliser ces termes pour grouper en catégories leurs activités de formation.

Les constatations faites à cette étape initiale de l'essai ont montré l'importance de consulter les répondants avant de fixer définitivement la terminologie et les concepts utilisés dans les questionnaires. Les constatations de l'étude ont fourni à l'équipe de projet de l'enquête des idées et des renseignements importants sur la façon de formuler les questions de l'enquête et de grouper en catégories les options de réponses.

Par exemple, une constatation importante découlant des réunions des groupes de discussion et des interviews en profondeur était que, dans de nombreuses sociétés, on n'utilisait pas les termes "régulière" ou "non structurée" pour décrire les activités de formation et que ces sociétés ne voyaient ni l'avantage ni la nécessité de faire cette distinction. Beaucoup de répondants ne voyaient pas non plus, entre les termes "régulière" et "non structurée", de nette distinction qui faciliterait le groupement des activités de formation en catégories.

L'étude a aidé les concepteurs de l'enquête à comprendre comment les répondants interprétaient les termes et les concepts. Les participants ont fait des suggestions sur la terminologie qui leur semblait appropriée. Par exemple, même s'ils éprouvaient des difficultés à propos des termes "régulière" et "non structurée", les participants ont pu proposer des caractéristiques pour définir ces termes. Ils ont dit de la formation régulière qu'elle était un programme d'études ou un plan de cours structuré avec un début, un milieu et une fin, avec des objectifs connus ou des buts bien définis; qu'elle comportait une composante d'évaluation;

et qu'elle avait un coût. D'autre part, la plupart des participants considéraient la "formation non structurée" comme une formation en cours d'emploi sans structure, souvent comme une forme d'apprentissage par l'observation. L'"absence d'évaluation" était une autre caractéristique souvent proposée pour définir la formation non structurée. Une autre constatation intéressante est que beaucoup de participants faisaient une distinction entre "formation" et "activités de perfectionnement ou éducatives". On ne considérait pas que le terme "formation" englobait toutes les activités des employés pour aider au perfectionnement de leur personnel. Certains participants considéraient la "formation" comme propre au poste et liée à la productivité du travail et le "perfectionnement" comme relatif à l'acroissement de la base de connaissances de la personne (Kennedy et de Groh 1992).

Après l'élaboration de l'ébauche du questionnaire de l'ENF, on en a fait l'essai au moyen de groupes de discussion et d'interviews simultanées où les répondants sont invités à livrer spontanément leur état d'esprit. Les représentants de diverses entreprises et d'un ensemble de petites, moyennes et grandes sociétés ont participé à l'étude. Les questions suivantes ont été étudiées:

- La personne qui, dans une entreprise, était la plus compétente pour répondre à l'enquête.
- La meilleure façon d'atteindre les répondants.
- Le processus suivi par les répondants pour fournir les renseignements demandés.
- La façon dont les répondants comprenaient les questions et les instructions.
- La réaction des répondants au vocabulaire et ainsi qu'aux groupements et classifications des professions dans l'enquête.
- Si les renseignements recherchés dans le cadre de l'enquête étaient faciles à obtenir.
- Les genres de dossiers dont étaient tirés les renseignements.
- La comparabilité des questions et des catégories de réponses avec les pratiques des répondants en matière de tenue des dossiers.
- Si les périodes de référence mentionnées dans l'enquête correspondaient aux pratiques des répondants en matière de tenue des dossiers.
- Le fardeau de déclaration du point de vue du temps et de l'effort.

À Ottawa, Toronto, Montréal et Vancouver, on a réuni en tout 7 groupes de discussion et on a réalisé 26 interviews. Dans le rapport final (D.R. Harley Consultants Limited 1993), l'entrepreneur a signalé de nombreuses constatations et fait plusieurs recommandations susceptibles d'améliorer le questionnaire.

Comme dans d'autres études sur les enquêtes-entreprises, une constatation importante était que beaucoup de participants mettaient en doute l'objet de l'enquête. Ils voulaient savoir pourquoi les renseignements étaient recueillis et à quoi allaient servir les résultats de l'enquête. Un thème important qui est ressorti de toutes les rencontres de groupes de discussion et des interviews était que les répondants voulaient savoir en quoi l'enquête leur serait utile.

Figure 2 (après l'essai): Enquête sur l'industrie de la construction de 1989 (Entrepreneurs généraux et promoteurs), Statistique Canada

SECTION 2. ÉTAT DES REVENUS

201

Dollars (Cmètre les cents)

2.1 Recettes au titre des contrats de construction

2.2 Autres recettes d'exploitation, notamment au titre de la vente de matériaux et de terrains, de la gestion de projets et de travaux de construction, de la location d'équipement ou d'immeubles et de la prestation de services de déneigement et d'experts-conseils en génie, etc. Veuillez préciser:

Description

2.3 Total des recettes brutes d'exploitation (total des sommes indiquées aux postes 202 et 207 à 210)

211

210

209

208

207

206

2.4 Veuillez indiquer la méthode comptable utilisée:

1 ☐ comptabilisation du revenu à l'achèvement des travaux

2 ☐ comptabilisation proportionnelle du revenu

212

FRAIS DIRECTS

2.5 Travaux en cours au début de l'exercice comptable (additionnez, si nécessaire, pour calculer les frais directs). On n'ont pas encore été facturés

2.6 Travaux en sous-traitance (y compris la location d'équipement avec opérateur)

2.7 Location d'équipement sans opérateur

2.8 Matériaux et fournitures utilisés (corriger de la variation du stock)

2.9 Rémunération brute versée à tous les ouvriers selon un taux horaire (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)

2.10 Traitements directs bruts imputés aux contrats et versés aux employés permanents, notamment aux contremaîtres, aux surveillants, etc. (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)

2.11 Contribution de l'employeur aux avantages sociaux versés aux employés, notamment les versements au titre des régimes de retraite et des assurances. (Déclarez ces frais **seulement** si les avantages sociaux versés aux employés ne sont pas déjà inclus dans la rémunération et les traitements directs indiqués ci-dessus.)

2.12 Coût des terrains inclus dans les ventes

2.13 Dépenses au titre de la réparation et de l'entretien des machines et de l'équipement

2.14 Frais d'amortissement imputables aux contrats

2.15 Autres frais directs (tous les autres frais directs qui ne sont pas déjà indiqués ci-dessus, par exemple les frais préalables à la construction, les coûts sur le chantier, les honoraires, les frais de publicité, de carburant, etc.)

2.16 Total des frais directs (total des chiffres indiqués aux postes 224 à 233)

2.17 Travaux en cours, à la fin de l'exercice comptable (déduire si nécessaire pour calculer les frais directs). La définition de travaux en cours figure à la question 2.5

2.18 Total des frais directs imputables aux contrats (poste 213 plus poste 234 moins poste 235)

SECTION 4. MAIN-D'OEUVRE

4.1 Veuillez déclarer le nombre d'heures travaillées par la main-d'oeuvre rémunérée au taux horaire (et dont les salaires ont été déclarés au poste 227):

N.B.: Les heures déclarées doivent correspondre aux heures travaillées, c.-à-d. qu'une heure supplémentaire payée à temps et demi équivaut à une heure. Les heures travaillées sont indiquées dans les dossiers des listes de paye ou dans les rapports de la Commission des accidents du travail.

401 heures

402 \$ par heure

4.2 Veuillez indiquer le nombre annuel moyen d'employés auxquels sont versés des traitements directs (et dont les traitements ont été déclarés aux postes 228):

403 employés

4.3 Veuillez indiquer le nombre annuel moyen d'employés généraux et administratifs auxquels sont versés des traitements bruts (et dont les traitements ont été déclarés au poste 237):

404 employés

4.4 Nombre d'ingénieurs inclus dans le chiffre inscrit aux postes 404:

405 ingénieurs

Figure 1 (avant l'essai): Recensement de l'industrie de la construction, 1988 (Entrepreneurs généraux et promoteurs), Statistique Canada

SECTION 2. ÉTATS DES RÉSULTATS																																									
RECETTES	<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>2.1 Recettes au titre des contrats de construction</p> <p>2.2 Autres recettes d'exploitation (précisez):</p> </div> <div style="width: 50%;"> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 10%;">102</td> <td style="width: 40%;"></td> <td style="width: 10%; text-align: right;">103</td> <td style="width: 10%; text-align: right;">\$</td> <td style="width: 10%;"></td> <td style="width: 10%;"></td> </tr> <tr> <td>104</td> <td></td> <td style="text-align: right;">105</td> <td></td> <td></td> <td></td> </tr> <tr> <td>106</td> <td></td> <td style="text-align: right;">107</td> <td></td> <td></td> <td></td> </tr> <tr> <td>108</td> <td></td> <td style="text-align: right;">109</td> <td></td> <td></td> <td></td> </tr> <tr> <td colspan="2" style="text-align: right;">Total</td> <td></td> <td></td> <td></td> <td></td> </tr> </table> </div> </div>	102		103	\$			104		105				106		107				108		109				Total						101	110	111							
102		103	\$																																						
104		105																																							
106		107																																							
108		109																																							
Total																																									
<p>2.3 Total des recettes brutes d'exploitation (somme des postes 2.1 et 2.2)</p>																																									
<p>2.4 Méthode comptable utilisée: <input type="checkbox"/> 1 Comptabilisation du revenu à l'achèvement des travaux <input type="checkbox"/> 2 Comptabilisation proportionnelle du revenu</p>																																									
<p>FRAIS DIRECTS</p>																																									
<p>2.5 Travaux en cours au début de l'exercice comptable (additionner si nécessaire, pour calculer les frais directs</p>																																									
<p>À défaut de la ventilation des frais directs, fournissez les totaux en pourcentage (poste 2.15; la somme doit être égale à 100).</p>																																									
<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 15%;">2.6 Travaux en sous-traitance</td> <td style="width: 15%;"></td> <td style="width: 15%;">2.7 Matériaux et fournitures utilisés (corrige de la variation du stock)</td> <td style="width: 15%;"></td> <td style="width: 15%;">2.8 Rémunération brute des ouvriers à taux horaire (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)</td> <td style="width: 15%;"></td> </tr> <tr> <td>2.9 Traitements directs bruts versés aux surveillants, etc. (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)</td> <td></td> <td>2.10 Avantages sociaux (contributions de l'employeur non déclarées aux postes 2.8 et 2.9, notamment les versements au titre des régimes de retraite, de l'assurance, etc.)</td> <td></td> <td>2.11 Terrains</td> <td></td> </tr> <tr> <td colspan="3"></td> <td colspan="3"> <div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>Le coût comprend (cocher):</p> <p><input type="checkbox"/> 1 Propriété non bâtie</p> <p><input type="checkbox"/> 2 Terrain plus frais de service, de crédit, etc.</p> <p><input type="checkbox"/> 3 Terrain viabilisé</p> </div> <div style="width: 50%;"> <p>2.12 Dépenses au titre de la réparation et de l'entretien des machines et de l'équipement</p> <p>2.13 Location d'équipement (sans opérateur)</p> <p>2.14 Autres frais directs</p> </div> </div> </td> </tr> <tr> <td colspan="3"></td> <td colspan="3"> <p>2.15 Total des frais directs (somme des postes 2.6 à 2.14)</p> </td> </tr> <tr> <td colspan="6" style="padding: 5px;"> <p>2.16 Travaux en cours à la fin de l'exercice comptable (déduire si nécessaire, pour calculer les frais directs)</p> </td> </tr> <tr> <td colspan="6" style="padding: 5px;"> <p>2.17 Total des frais directs imputables aux contrats (poste 2.5 plus poste 2.15 moins 2.16)</p> </td> </tr> </table>						2.6 Travaux en sous-traitance		2.7 Matériaux et fournitures utilisés (corrige de la variation du stock)		2.8 Rémunération brute des ouvriers à taux horaire (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)		2.9 Traitements directs bruts versés aux surveillants, etc. (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)		2.10 Avantages sociaux (contributions de l'employeur non déclarées aux postes 2.8 et 2.9, notamment les versements au titre des régimes de retraite, de l'assurance, etc.)		2.11 Terrains					<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>Le coût comprend (cocher):</p> <p><input type="checkbox"/> 1 Propriété non bâtie</p> <p><input type="checkbox"/> 2 Terrain plus frais de service, de crédit, etc.</p> <p><input type="checkbox"/> 3 Terrain viabilisé</p> </div> <div style="width: 50%;"> <p>2.12 Dépenses au titre de la réparation et de l'entretien des machines et de l'équipement</p> <p>2.13 Location d'équipement (sans opérateur)</p> <p>2.14 Autres frais directs</p> </div> </div>						<p>2.15 Total des frais directs (somme des postes 2.6 à 2.14)</p>			<p>2.16 Travaux en cours à la fin de l'exercice comptable (déduire si nécessaire, pour calculer les frais directs)</p>						<p>2.17 Total des frais directs imputables aux contrats (poste 2.5 plus poste 2.15 moins 2.16)</p>					
2.6 Travaux en sous-traitance		2.7 Matériaux et fournitures utilisés (corrige de la variation du stock)		2.8 Rémunération brute des ouvriers à taux horaire (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)																																					
2.9 Traitements directs bruts versés aux surveillants, etc. (avant les retenues de l'impôt sur le revenu, des régimes de retraite, de l'assurance, etc.)		2.10 Avantages sociaux (contributions de l'employeur non déclarées aux postes 2.8 et 2.9, notamment les versements au titre des régimes de retraite, de l'assurance, etc.)		2.11 Terrains																																					
			<div style="display: flex; justify-content: space-between;"> <div style="width: 45%;"> <p>Le coût comprend (cocher):</p> <p><input type="checkbox"/> 1 Propriété non bâtie</p> <p><input type="checkbox"/> 2 Terrain plus frais de service, de crédit, etc.</p> <p><input type="checkbox"/> 3 Terrain viabilisé</p> </div> <div style="width: 50%;"> <p>2.12 Dépenses au titre de la réparation et de l'entretien des machines et de l'équipement</p> <p>2.13 Location d'équipement (sans opérateur)</p> <p>2.14 Autres frais directs</p> </div> </div>																																						
			<p>2.15 Total des frais directs (somme des postes 2.6 à 2.14)</p>																																						
<p>2.16 Travaux en cours à la fin de l'exercice comptable (déduire si nécessaire, pour calculer les frais directs)</p>																																									
<p>2.17 Total des frais directs imputables aux contrats (poste 2.5 plus poste 2.15 moins 2.16)</p>																																									

SECTION 4. MAIN-D'OEUVRE					
4.1 En ce qui concerne le poste 2.8, soit la rémunération versée aux ouvriers à taux horaire, déclarez le nombre d'heures travaillées:	201	heures	ou taux horaire moyen: \$	202	/ heure
<p>N.B.: Les heures déclarées doivent être des heures travaillées, c.-à-d. qu'une heure supplémentaire payée à temps et demi équivaut à une heure.</p>					
<p>4.2 En ce qui concerne le poste 2.9, traitements directs versés, indiquez le nombre annuel moyen d'employés:</p>					
<p>4.3 En ce qui concerne le poste 2.19, traitements bruts versés, déclarez le nombre annuel moyen d'employés:</p>					

Section 2 – État des résultats

Sur la version finale du questionnaire (figure 2):

- Une déclaration figure au début de la section 2 dans laquelle on informe les répondants qu'ils peuvent inclure les états financiers de leur société. Sur la version du questionnaire (figure 1) qui a fait l'objet de l'essai, de nombreux répondants n'ont pas vu cette instruction parce qu'elle figurait sur une page d'instructions distincte.
- On se réfère aux numéros de ligne (p. ex. 202 et 207-210) plutôt qu'aux numéros de poste (p. ex. 2.1 et 2.2). Bien que les numéros de ligne soient en fait des numéros de code pour les données, les répondants les considéraient comme des numéros de ligne parce qu'ils ressemblaient à l'usage commun bien connu de numéros de ligne sur les formulaires utilisés pour les déclarations d'impôt au Canada.
- Les renseignements importants tels que les définitions et l'indication des renseignements à donner figurent dans les postes eux-mêmes plutôt que sur la pagée d'instructions. C'est seulement dans les cas où les détails relatifs aux coûts directs ne sont pas disponibles que les répondants doivent déclarer des pourcentages estimés. On a rendu ce choix plus évident en utilisant des caractères gras et plus gros pour le mot "ou".

À noter que pour remplir la section 2 les répondants ont consulté les sources de renseignements suivantes: états financiers, systèmes de comptabilité en direct, facturation proportionnelle et facturation d'ouvrages en cours, comptes rendus de projets, grands livres généraux, documents de travail et certificats de vérification.

Section 4 – Main-d'œuvre

Sur la version finale du questionnaire (figure 2):

- À la question 4.1, on précise que les "heures travaillées" peuvent être tirées des "dossiers des listes de paye ou des rapports pour les commissions des accidents du travail". Pendant les interviews où les répondants sont invités à livrer spontanément leur état d'esprit, les répondants ont fait remarquer qu'ils avaient consulté des dossiers de ce genre pour obtenir les renseignements demandés.
- On précise que le "taux horaire moyen" doit être déclaré "À défaut des heures travaillées".
- Des instructions et des renseignements importants accompagnent les questions. Par exemple, pendant l'essai, la plupart des répondants n'ont pas exclu les propriétaires et les associés quand ils ont déclaré le nombre d'employés aux questions 4.2 et 4.3 (même si cette exclusion était précisée sur la page d'instructions).

5.2 Enquête nationale sur la formation (ENF)

Deux études distinctes, chacune comportant l'utilisation de groupes de discussion et de méthodes de recherche cognitive, ont été effectuées pendant l'élaboration et l'essai du questionnaire de l'enquête nationale sur la formation (ENF).

questionnaires et formulaires qu'ils devaient remplir chaque année. Un bon nombre des participants ont précisé qu'ils attendaient souvent l'appel téléphonique de suivi avant de répondre et certains préféraient même répondre au téléphone. Ils ont dit qu'au téléphone ils pouvaient faire des estimations au juge plutôt que de remplir le questionnaire avec soin et que cela leur demandait beaucoup moins de temps et d'effort.

Le fardeau de déclaration était plus une impression qu'une réalité. Après avoir rempli le questionnaire, beaucoup de répondants ont fait remarquer qu'ils avaient consacré beaucoup moins de temps qu'ils ne l'avaient prévu à remplir le questionnaire et que ce travail avait été plus facile qu'ils ne s'y étaient attendus.

Un thème commun qui est ressorti des interviews et des rencontres des groupes de discussion était la valeur accordée par les répondants aux renseignements demandés. Les répondants voulaient savoir pourquoi ils devaient remplir le questionnaire et souvent mettaient en doute la valeur des renseignements pour eux-mêmes et pour les autres utilisateurs des renseignements. Une des choses importantes que la recherche a permis de constater est donc l'importance de bien faire comprendre aux répondants la valeur des renseignements qu'ils fournissent. Ces répondants voulaient savoir à quoi serviraient les résultats de l'enquête. Ils voulaient aussi savoir comment obtenir les données.

Globalement, les questionnaires ont été très bien accueillis par les répondants. Ces derniers en ont apprécié l'apparence et l'approche "professionnelles". Beaucoup avaient déjà rempli les questionnaires de précédents recensements de l'industrie de la construction. Ils considéraient que les questionnaires remanés étaient une amélioration sur les versions antérieures parce qu'ils semblaient moins longs et moins compliqués. Cela était une réaction positive et rassurante pour les gestionnaires de l'enquête qui avaient conçu les nouveaux questionnaires (Gower et Zylstra 1990; Price Waterhouse Conseillers en gestion 1990).

L'étude a permis de tirer de nombreuses conclusions précises à propos de la façon dont les questionnaires pouvaient être améliorés et rendus plus conviviaux pour les répondants. L'essai préliminaire a permis d'obtenir une rétroaction intéressante à propos des taux de réponse et de la complétude de la déclaration, et les groupes de discussion et la recherche cognitive ont ajouté beaucoup à ces conclusions en fournissant une information approfondie et de première main sur la façon dont les répondants réagissent aux questions et la raison pour laquelle ils le font et sur la façon dont les réponses ont été choisies et la raison pour laquelle elles l'ont été.

Les figures 1 et 2 montrent comment le questionnaire a été amélioré d'après quelques-unes des constatations particulières qui ont été faites (Gower 1993). La figure 1 montre des parties des sections 2 et 4 de la version de 1988 du questionnaire utilisé pour les entrepreneurs généraux et promoteurs, avant l'essai. La figure 2 montre les parties correspondantes de la version finale de ce questionnaire, après l'essai.

- Enquête sur le commerce de gros et de détail (Noonan 1992).
- Enquête nationale sur la formation (Kennedy et de Groh Consultants 1992; D.R. Harley Consultants Limited 1993).

Ces études comportaient l'application d'au moins une des méthodes suivantes: groupes de discussion, interviews en profondeur, interviews simultanées où les répondants sont invités à livrer spontanément leur état d'esprit, reformulation. Toutes les études ont été réalisées sous la coordination et la direction générale du Centre d'information sur la conception des questionnaires de Statistique Canada (Gower 1991).

Toutes ces études ont montré l'avantage et l'importance de consulter les membres de la population cible avant d'élaborer le questionnaire et d'en préparer la version finale. Elles ont révélé des aspects très intéressants du processus de réponse et divers facteurs d'erreur de mesure dans les enquêtes-entreprises. Ces facteurs sont la valeur que les répondants accordent aux renseignements demandés, leur perception du fardeau de déclaration, la compatibilité des questions avec les pratiques en matière de tenue des dossiers, la place et l'utilisation des instructions, la disponibilité des données et la complexité du processus de réponse (Gower et Zylistra 1990).

On traite ci-après d'aspects importants de deux de ces études, le recensement de l'industrie de la construction et l'enquête nationale sur la formation.

5.1 Recensement de l'industrie de la construction

Le recensement annuel de l'industrie de la construction a été conçu pour recueillir des statistiques complètes sur l'industrie de la construction au Canada. La population cible était composée des établissements dont le revenu principal provenait d'une activité liée à la construction. Il y avait deux questionnaires distincts: un pour les entrepreneurs généraux et promoteurs et un second pour les entrepreneurs de métiers spécialisés et sous-traitants. Les questionnaires postés aux répondants permettaient de recueillir des données sur les revenus et les dépenses, sur le travail et sur la distribution des extrants.

Les questionnaires utilisés en 1988 pour le recensement de l'industrie de la construction ont été remaniés pour l'enquête de 1989. Les principaux objectifs de la révision étaient de réduire le contenu des questionnaires et le fardeau de déclaration et de répondre à la nécessité d'apporter des améliorations importantes aux questionnaires existants.

Un essai préliminaire des questionnaires révisés a été réalisé pour obtenir la réaction des entrepreneurs (Statistique Canada 1989). L'essai préliminaire a montré que les questionnaires révisés étaient bien reçus et compris par les répondants. On a relevé certains domaines où d'autres améliorations devaient être apportées, par exemple la formulation des questions et la clarté de certaines instructions.

Pour obtenir plus de renseignements sur la façon dont les répondants considéreraient les questionnaires révisés et pour s'assurer que les taux de réponse et la qualité des données seraient maximisés, d'autres essais des questionnaires ont

été réalisés au début de 1990, à l'aide de groupes de discussion et de méthodes cognitives. Cette phase de l'essai a été conçue pour obtenir une information approfondie sur les points suivants:

- L'impression des répondants quant aux questionnaires. Le processus que les répondants suivaient pour fournir les renseignements.
- L'agencement, la présentation et la lisibilité des questionnaires.
- La mesure dans laquelle les répondants lisaient et comprennent les instructions et les questions.
- Les problèmes rencontrés par les répondants lorsqu'ils remplissaient les questionnaires.
- Si les instructions et les définitions étaient nécessaires, compréhensibles et utiles.
- L'exactitude des renseignements fournis par les répondants.
- L'utilisation d'estimations par les répondants et l'exactitude de ces estimations.
- Les types de dossiers dont les renseignements ont été tirés. La compatibilité des questions et des catégories de réponses avec les pratiques des répondants en matière de tenue de dossiers.
- Le fardeau de déclaration du point de vue du temps et de l'effort.

La recherche englobait le questionnaire sur les entrepreneurs généraux et promoteurs et le questionnaire sur les entreprises de métiers spécialisés et sous-traitants. Environ 50 entreprises de construction ont participé à l'étude. Elles ont été choisies de manière à représenter les genres de répondants qui remplissaient les questionnaires du recensement de l'industrie de la construction. On a réalisé à Ottawa, Montréal et Toronto 25 interviews en profondeur et 16 interviews simultanées où les répondants sont invités à livrer spontanément leur état d'esprit et on y a tenu des réunions pour deux groupes de discussion. Toutes les interviews individuelles ont eu lieu à l'établissement du répondant.

Une conclusion très intéressante de l'étude est qu'il y avait deux groupes distincts de répondants. Le premier groupe était composé du président ou du vice-président d'une société, qui souvent devait consulter d'autres personnes pour répondre à certaines questions. Ces participants ont pris de 35 à 45 minutes pour remplir le questionnaire. Ils étaient plus susceptibles de faire des estimations d'après leur connaissance de l'entreprise et moins susceptibles d'expliquer les écarts entre le questionnaire et la source de renseignements utilisée pour y répondre.

D'autre part, les chefs de bureau, les comptables et les contrôleurs ont pris de 75 à 90 minutes pour remplir le questionnaire. Ces répondants étaient beaucoup plus préoccupés par les détails et soucieux de fournir des réponses exactes. Ils utilisaient plusieurs sources de renseignements et faisaient des calculs pour pouvoir répondre aux questions (Gower et Zylistra 1990; Gower et Nargundkar 1991).

De nombreux répondants ont précisé que répondre au questionnaire n'était pas une priorité pour eux. Ils considéraient le questionnaire comme un parmi de nombreux

4.8 Méthodes d'essai formel

Les méthodes d'essai formel sont de nature quantitative. Elles sont conçues pour permettre une évaluation statistique de la façon dont le questionnaire atteint le but visé. Les études pilotes et l'essai avec échantillon fractionné sont deux genres de méthodes d'essai formel utilisées couramment. Ces méthodes conviennent mieux pour des enquêtes à grande échelle et permanentes parce que les essais et l'analyse des résultats sont plus coûteux.

On effectue une *étude pilote* pour observer comment les opérations liées à l'enquête, y compris le fait de faire remplir le questionnaire, fonctionnent ensemble en pratique. Une étude pilote est une "répétition générale". Elle reprend exactement le plan final de l'enquête sur une échelle restreinte, du début à la fin, y compris le dépouillement et l'analyse des données. Elle permet à la personne qui effectue la recherche relative à l'enquête de voir dans quelle mesure le questionnaire s'intègre bien aux autres parties de l'enquête. Certains problèmes peuvent être relevés seulement par un essai de toutes les phases de l'enquête ensemble. Par exemple, il se peut qu'on puisse, pendant la formation des intervieweurs, relever les fautes de frappe, les défauts de formulation d'une question ou le manque de précision d'un concept. L'étape du traitement informatique peut faire ressortir des fautes de frappe dans les numéros de postes codés à l'avance ou dans les catégories de réponses (DeMaio 1983).

Normalement, le questionnaire doit faire l'objet d'un essai préliminaire complet avant la réalisation d'une enquête pilote. L'enquête pilote n'est habituellement pas l'occasion d'essayer de nouvelles questions ou de nouvelles méthodes. Si l'on a déjà effectué un essai, il est peu probable que l'enquête pilote entraînera des modifications importantes dans le questionnaire. L'étude pilote donne toutefois l'occasion d'apporter les dernières retouches au questionnaire avant son utilisation dans l'enquête principale (DeMaio 1983).

On effectue un *essai avec échantillon fractionné* pour déterminer laquelle de deux ou de plusieurs versions du questionnaire est la "meilleure". L'essai avec échantillon fractionné est aussi appelé expérience "avec questionnaire fractionné" ou à "panel fractionné". Pour un essai de ce genre, on a recours à un plan d'expérience incorporé au processus de collecte des données. Un essai avec échantillon fractionné peut être conçu pour étudier des problèmes tels que la formulation et l'enchaînement des questions, les endroits du questionnaire où sont placées les questions, délicates et les procédures de collecte des données. Dans un plan simple avec échantillon fractionné, une moitié de l'échantillon est choisie au hasard et peut recevoir un traitement expérimental tandis que l'autre moitié reçoit l'autre traitement. Dans un essai qui comporte deux traitements expérimentaux, on pourrait avoir recours à un plan factoriel 2×2 où l'on fait l'essai de chacun des deux traitements dans chaque expérience sur une moitié de l'échantillon (DeMaio 1983).

On peut aussi utiliser un plan avec échantillon fractionné dans des enquêtes permanentes qui évaluent les tendances dans le temps et qui comparent les résultats d'une enquête

5. GROUPES DE DISCUSSION ET MÉTHODES

DE RECHERCHE COGNITIVE POUR L'ESSAI DES QUESTIONNAIRES D'ENQUÊTES-ENTREPRISES

Statistique Canada a constaté que les groupes de discussion et les méthodes de recherche cognitive sont très utiles pour élaborer et mettre à l'essai les questionnaires d'enquêtes-entreprises. Ces méthodes donnent la possibilité de comprendre les processus cognitifs qui entrent en jeu lorsqu'on formule les réponses aux questions d'une enquête. Elles introduisent la perspective du répondant directement dans le processus de conception du questionnaire et permettent de concevoir des questionnaires conviviaux pour les répondants (Gower et Nargundkar 1991). Statistique Canada a utilisé les groupes de discussion et les méthodes de recherche cognitive dans le domaine des enquêtes-entreprises pour élaborer et mettre à l'essai les questionnaires des enquêtes suivantes:

- Enquête sur l'emploi la rémunération et les heures de travail (Bureau 1991; Goss, Gilroy et Associates Ltd. 1989; Goss, Gilroy et Associates Ltd. 1990).
- Recensement de l'industrie de la construction (Gower et Zyisira 1990; Price Waterhouse Consultants en gestion 1990).

4.9 Examen et révision du questionnaire

Le questionnaire doit être examiné par une personne qui ne fait pas partie de l'équipe de projet. Les examinateurs peuvent être des spécialistes ou des personnes qui ont de l'expérience dans la conception de questionnaires. Un examen effectué à toutes les étapes du processus d'élaboration du questionnaire ou à n'importe laquelle d'entre elles peut entraîner des révisions aux questions et aux catégories de réponses.

La conception d'un questionnaire est un processus itératif. Pendant tout le processus d'élaboration, de révision et d'essai, des changements seront apportés continuellement pour améliorer le questionnaire. Les objectifs et les besoins en renseignements sont précisés, évalués et les utilisateurs des données et les répondants sont consultés, les questions proposées sont rédigées et mises à l'essai, les questions sont examinées et révisées jusqu'à ce qu'un questionnaire final soit établi.

exprimer clairement et spontanément son état d'esprit. Il arrive parfois que l'intervieweur doive aider le répondant dans cette tâche en lui posant des questions simples: "à quelle question répondez-vous actuellement?", "à quel point pensez-vous actuellement?", "veuillez expliquer comment vous avez choisi la réponse", ou d'autres questions visant à préciser les pensées du répondant. Quand un répondant est peu disposé à exprimer son état d'esprit, l'observateur peut décider qu'il est préférable de traiter l'interview comme une interview en profondeur et agir en conséquence.

Les interviews où les répondants sont invités à livrer spontanément leur état d'esprit sont très utiles pour obtenir les réactions des répondants relativement à des questionnaires. Elles sont particulièrement utiles pour déterminer les parties du questionnaire où les répondants éprouvent des difficultés. Elles aident aussi le chercheur à comprendre le processus suivi pour répondre au questionnaire.

- *Groupe de discussion*: Comme on l'a vu dans la partie 4.4, les groupes de discussion sont utilisés pour évaluer dans quelle mesure les répondants comprennent le langage et la formulation des questions et des instructions. Les personnes qui font partie du groupe de discussion doivent habituellement, avant la réunion du groupe, soit répondre au questionnaire au cours d'une interview sur place ou au téléphone, soit remplir elles-mêmes le questionnaire. Pendant la réunion du groupe de discussion, l'animateur passe le questionnaire en revue avec les participants et discute de tout problème ou de toute difficulté que ces derniers peuvent avoir rencontré lorsqu'ils répondaient au questionnaire. Les groupes de discussion stimulent et encouragent l'analyse réfléchie du questionnaire pendant les discussions de groupe sur les commentaires de chacun des participants. Ils sont particulièrement utiles pour susciter des suggestions et des recommandations visant à améliorer le questionnaire.

- *Reformulation*: La reformulation est utilisée dans les interviews individuelles et dans les groupes de discussion. On demande aux répondants de répéter la question dans leurs propres mots ou d'expliquer la signification des termes et concepts utilisés dans les questions et instructions du questionnaire. La reformulation aide à déterminer si les répondants lisent et comprennent correctement les instructions et les questions. Elle est particulièrement utile pour trouver les libellés de questions trop complexes ou prêtant à confusion. Elle permet aussi de découvrir des situations où les répondants ne comprennent pas tous les éléments importants de la question (p. ex. la période de référence).

4.7 Essai préliminaire

L'*essai préliminaire* est une étape fondamentale de l'élaboration d'un questionnaire. Il consiste habituellement à effectuer un petit nombre d'interviews sur le terrain pour relever les problèmes liés à un questionnaire. L'essai peut porter sur tout le questionnaire ou sur une partie seulement.

Les essais préliminaires permettent de déceler des formulations ou des enchaînements de questions qui laissent à désirer, des erreurs dans la présentation des questionnaires ou dans les instructions et des problèmes causés par l'inaptitude ou la réticence du répondant à répondre aux questions. Les essais préliminaires sont aussi utilisés pour suggérer des catégories de réponses additionnelles qui peuvent être codées à l'avance sur le questionnaire. Ces essais donnent une idée de la durée des interviews et des problèmes liés aux refus.

L'échantillon de l'essai préliminaire peut comprendre de 20 à 100 répondants ou même plus. Si le but principal de l'essai préliminaire est de trouver les problèmes liés à la formulation ou à l'enchaînement des questions, il se peut que seulement un petit nombre d'interviews soient nécessaires. On a besoin de plus d'interviews (de 50 à 100) pour déterminer les catégories de réponses codées à l'avance dans le cas des questions à réponse libre. Les répondants aux essais préliminaires sont habituellement choisis à dessein plutôt qu'au hasard.

La méthode utilisée pour répondre au questionnaire dans le cadre d'un essai préliminaire doit être la même que celle qui est prévue pour l'enquête principale (p. ex. questions posées par l'intervieweur sur place ou par téléphone). L'essai préliminaire d'un questionnaire postal est plus efficace si l'on a recours à des intervieweurs. On peut employer ces derniers pour livrer le questionnaire et, plus tard, pour discuter de tout problème rencontré par le répondant. Les concepteurs du questionnaire doivent observer le plus grand nombre possible d'interviews réalisées lors de l'essai préliminaire.

L'essai préliminaire n'est pas aussi efficace que les méthodes cognitives pour évaluer comment les répondants comprennent une question et la difficulté du processus de réponse. L'essai préliminaire ne fait que montrer s'il y a un problème. Sans autre examen, il ne permet de trouver ni la cause ni la solution du problème.

On tient souvent des *séances de compte rendu* avec les intervieweurs conjointement avec un essai préliminaire. Les intervieweurs qui ont participé à un essai préliminaire peuvent signaler des points importants sur lesquels on peut améliorer le questionnaire. Quand il s'agit de questionnaires à remanier, il est utile de consulter les intervieweurs de manière qu'il puisse être tenu compte de leur opinion dans le processus de remaniement. Les intervieweurs ont d'excellentes idées sur la manière de faire remplir un questionnaire et sur son effet sur la collaboration des répondants.

On peut aussi effectuer un *codage du comportement* au moment de l'essai préliminaire. On effectue un enregistrement sonore de l'interview, après quoi on code et analyse le comportement qu'ont eu l'intervieweur et le répondant pendant l'interview. Le codage du comportement est un moyen systématique et objectif d'examiner l'efficacité du questionnaire. Il aide aussi à déterminer quels domaines posent des problèmes, par exemple le fait, pour l'intervieweur, de ne pas lire la question telle qu'elle est formulée ou, pour le répondant, de demander de préciser la question ou le processus de réponse.

- Utilisez les **caractères gras** ou le soulignement pour mettre l'accent sur des éléments importants comme la période de référence ou de déclaration.
- Précisez "incluez" ou "n'incluez pas" dans les questions ou les postes eux-mêmes (et non dans des instructions à part).

Voici d'autres considérations qui entrent en ligne de compte dans la conception des questionnaires d'enquêtes-entreprises:

- La terminologie, les questions et les catégories de réponses doivent être compatibles avec les concepts et définitions courants.
- Les caractéristiques des répondants, par exemple les pratiques en matière de tenue des dossiers et la compétence linguistique.
- La disponibilité des données.
- Le fardeau de déclaration.
- La complexité des données à recueillir.
- La comparabilité des résultats avec ceux d'autres enquêtes.
- La fiabilité des données.
- La non-réponse.

Dans la conception du questionnaire, il faut aussi tenir compte de toute exigence administrative de l'organisme qui réalise l'enquête. Par exemple, Statistique Canada a une politique d'information des répondants à une enquête (Statistique Canada 1986) qui exige que les renseignements essentiels soient expliqués aux répondants. Il faut leur expliquer le ou les buts principaux de l'enquête, les principales utilisations prévues des données, la nécessité de répondre (enquête obligatoire ou enquête à participation volontaire), la protection de la confidentialité et tout accord de collecte conjointe ou de partage des données. À Statistique Canada, il y a aussi d'autres exigences administratives ou juridiques. Par exemple, la Loi sur les langues officielles du Canada exige que les répondants puissent obtenir les questionnaires dans l'une ou l'autre langue officielle (c.-à-d. en anglais ou en français).

4.6 Utilisation de méthodes cognitives pour faire l'essai des questionnaires

L'essai des questionnaires est une étape essentielle de l'élaboration de questionnaires efficaces qui permettent de recueillir des données utiles et exactes. Les méthodes de recherche cognitive, parfois appelées essais qualitatifs, sont particulièrement utiles pour l'essai de questionnaires. Les méthodes cognitives fournissent le moyen d'observer la démarche de la pensée des répondants au moment où ils répondent aux questions d'une enquête. Elles sont utilisées pour vérifier si les répondants comprennent ou non ce que signifient les questions et aident à évaluer la validité des questions et à déterminer des causes possibles d'erreur de mesure. Les méthodes cognitives offrent aussi la possibilité d'évaluer le questionnaire du point de vue du répondant. Elles portent sur des points tels que la compréhension et les réactions au questionnaire. Cela fait entrer la perspective du répondant directement dans le processus de

conception du questionnaire. L'utilisation de méthodes cognitives mène à la conception de questionnaires conviviaux pour les répondants et qui puissent être remplis facilement et avec exactitude.

Dans les enquêtes-entreprises, les méthodes cognitives sont utilisées pour étudier le rapport entre le répondant et la source de renseignements externe. On les emploie aussi pour étudier l'influence de cette source sur le processus de réponse. Ces méthodes donnent le moyen d'évaluer la compatibilité de la formulation des questions, des périodes de référence et des catégories de réponses avec les parties des entreprises en matière de tenue des dossiers.

Les méthodes d'essai cognitif (Gower 1993) sont les suivantes:

- *Interviews en profondeur*: Cette technique comprend l'interview individuelle (séance rétrospective où les répondants sont invités à livrer spontanément leur état d'esprit). Dans le cas d'un questionnaire envoyé par la poste, les répondants remplissent d'abord le questionnaire comme ils le feraient normalement. Un intervieweur observe le processus, relevant l'ordre dans lequel l'enquête répond aux questions, les occasions où il se reporte aux instructions et les autres personnes ou les genres de dossiers consultés. L'intervieweur relève aussi le temps nécessaire pour répondre aux diverses sections et les corrections ou modifications apportées aux réponses.
- Puis l'intervieweur réalise l'interview en profondeur et obtient des renseignements à propos des impressions qu'a eues le répondant lorsqu'il remplissait le questionnaire et des expériences antérieures qu'il aurait eues à cet égard. La discussion de suivi comporte en général un examen, question par question, du questionnaire avec le répondant qui permet de passer en revue toute difficulté ou tout problème rencontré par le répondant pendant qu'il remplissait le questionnaire. L'intervieweur demande des précisions pour savoir comment les termes et concepts ont été interprétés par les répondants, comment et pourquoi ils ont choisi les réponses et comment ils se sont remémoré les renseignements demandés.

Dans le cas d'un questionnaire pour lequel les questions sont posées par un intervieweur, celui-ci commence par poser les questions soit sur place, soit au téléphone. La discussion de suivi en profondeur a lieu après cette première interview.

- *Interviews simultanées où les répondants sont invités à livrer spontanément leur état d'esprit*: Il s'agit aussi d'interviews individuelles. On demande au répondant de répondre spontanément son état d'esprit¹ au moment où il répond aux questions, de faire des commentaires sur chaque question et d'expliquer comment il a choisi la réponse donnée. L'observateur peut poser des questions supplémentaires à propos des réponses pour obtenir plus de renseignements sur une déclaration particulière ou préciser le processus suivi pour choisir une réponse.
- Le succès de l'interview simultanée où les répondants sont invités à livrer spontanément leur état d'esprit dépend de deux facteurs: le répondant peut-il et veut-il

Les groupes de discussion sont un moyen de consulter les répondants, les utilisateurs des données et les interveneurs. Dans les premières étapes de l'élaboration d'un questionnaire, on a recours aux groupes de discussion pour élaborer les objectifs de l'enquête et les besoins en données, déterminer des projets de recherche importants et préciser les définitions et les concepts.

Les groupes de discussion sont aussi utiles pour effectuer l'essai et l'évaluation de questionnaires (voir 4.6). On y a recours pour évaluer comment les répondants comprennent le langage et la formulation utilisés dans les questions et les instructions et pour évaluer d'autres formulations et présentations des questions.

Le recrutement de participants venant des entreprises soulève des difficultés particulières pour les groupes de discussion. Il se peut que les stimulants pécuniaires ou la rétribution habituellement offerts aux participants de groupes de discussion (actuellement de 30 à 50\$ par participant) ne soient pas appropriés dans le cas des gens d'affaires. Des assurances en matière de confidentialité et l'accent mis sur l'importance de l'enquête et de la participation de ces personnes à l'étude sont plus significatifs. Un autre genre de motivation qui peut être offert est un don à un organisme sans but lucratif choisi par le participant. Statistique Canada donne souvent aux membres de groupes de discussion un exemplaire d'une publication qui les intéresse.

La taille des groupes de discussion varie de six à douze personnes. La taille optimale est de sept ou huit personnes pour les participants venant du milieu des affaires, bien que l'on forme parfois des groupes moins nombreux, de quatre ou cinq personnes (mini-groupes). À cause de la difficulté de trouver des participants venant des entreprises, les réunions des groupes de discussion doivent avoir lieu à un moment convenant aux participants. Dans le cas des gens d'affaires, les réunions des groupes de discussion se tiennent souvent pendant les heures de travail. On réalise un enregistrement sonore des réunions des groupes de discussion et les participants sont observés par des personnes qui se trouvent derrière une glace sans tain dans une pièce voisine. On informe les participants que la réunion est enregistrée sur bande sonore et que des personnes les observent.

4.5 Considérations entrant en ligne de compte dans la préparation des questions

Il faut tenir compte de nombreux éléments pour la formulation des questions et l'élaboration des catégories de réponses. Il importe de ne pas oublier les objectifs et les besoins en données et la façon dont les renseignements seront recueillis et traités. Les questions doivent être élaborées en fonction des besoins en information. Elles doivent être adressées aux bonnes personnes dans l'orga-

La méthode de collecte des données déterminera comment les questions et les catégories de réponses seront formulées. Le libellé des questions doit être clair et les questions doivent s'enchaîner selon un ordre logique. Elles doivent être conçues de façon que les répondants puissent les

Le questionnaire doit avoir une apparence professionnelle et sérieuse. Au moment de la conception du questionnaire, il ne faut pas oublier que les entreprises ont à remplir beaucoup de formules et de questionnaires. Le fait de remplir ces formules et ces questionnaires ne constitue pas une priorité pour les entreprises. Voici, selon les résultats d'une recherche effectuée par le Centre d'information sur la conception des questionnaires, des réactions typiques d'entreprises relativement au questionnaire de ce genre:

- "Je commence par remplir le questionnaire le plus court."
- "Suis-je obligé de le remplir?"
- "Y a-t-il une date limite pour le retour du questionnaire?"

Dans une étude réalisée par Statistique Canada (Gower et Zylstra 1990), un répondant a déclaré que si la réponse aux deux dernières questions ci-dessus était "non", alors il placerait le questionnaire dans le panier où il met les documents dont il s'occupe *peut-être un jour*.

Il arrive souvent que les répondants mettent en doute la valeur des renseignements pour eux-mêmes et pour les autres utilisateurs. Certains répondants souhaitent recevoir une réaction à propos de l'enquête. Par conséquent:

- Expliquez pourquoi il importe de remplir le questionnaire. Assurez-vous que les répondants comprennent bien l'importance des renseignements demandés.
- Expliquez comment seront utilisées les données de l'enquête.
- Expliquez comment les répondants peuvent obtenir les données.

Il faut aussi accorder une attention particulière aux instructions qui accompagnent le questionnaire. Une recherche effectuée par le Centre d'information sur la conception des questionnaires a montré à maintes reprises que les répondants ne lisent que ce qu'ils *pensent* devoir lire. Ils comprennent par lire ce qui est en caractères gras puis décident s'ils doivent poursuivre leur lecture. Les répondants lisent rarement les instructions; habituellement, ils passent directement aux questions. Ils ne se reportent aux instructions que s'ils *pensent* avoir besoin d'aide. C'est pourquoi il se peut que les répondants ne prennent pas connaissance d'instructions et de définitions importantes. Souvent les erreurs de déclaration sont dues au manque d'instructions claires et au fait que les répondants ne les lisent ou ne les comprennent pas (p. ex. ce qu'il faut inclure ou exclure). Par conséquent:

- Assurez-vous que les instructions sont brèves et claires.
- Dites aux répondants où trouver les instructions.
- Incluez les définitions au début du questionnaire ou, au besoin, dans des questions particulières.

d'enquêtes. Dans une petite entreprise, le répondant est souvent le propriétaire ou le chef de bureau, qui peut manquer de temps ou avoir un horaire trop peu souple pour remplir des questionnaires.

La nature des renseignements que doivent fournir les répondants à des enquêtes-entreprises oblige en général ces derniers à utiliser des dossiers ou d'autres systèmes d'information. Souvent les questionnaires utilisent une terminologie technique ou professionnelle associée à la communication de données financières ou administratives.

Le caractère confidentiel et la nature délicate des renseignements recueillis à l'aide du questionnaire sont un autre sujet de préoccupation. Dans de nombreux cas, les entreprises s'inquiètent du fait qu'elles doivent fournir des renseignements financiers confidentiels qu'elles ne souhaitent pas révéler à leurs concurrents, aux administrations publiques ou à toute autre partie. Il faut donc donner des assurances en matière de confidentialité. Toutes les mesures nécessaires doivent être prises pour assurer la manipulation et la garde appropriées des données de telle manière que soit préservée la confidentialité des renseignements.

3. LE PROCESSUS DE RÉPONSE DANS LES ENQUÊTES-ENTREPRISES

Le modèle du processus de réponse est bien connu pour les enquêtes-ménages. Pour répondre aux questions posées dans les enquêtes de ce genre, il faut faire appel à la compréhension, à la remémoration, à la pensée/appréciation et, finalement, répondre (Tourangeau 1984). Les répondants doivent d'abord comprendre la question. Ils cherchent ensuite dans leur mémoire pour extraire le renseignement demandé. Après s'être remémoré l'information, ils pensent à ce que pourrait être la réponse appropriée à la question et déterminent quelle partie de la réponse ils acceptent de donner. Ce n'est qu'à ce moment qu'ils donnent une réponse à la question.

Un modèle de réponse correspondant a aussi été élaboré pour les enquêtes-entreprises (Edwards et Cantor 1991). Bien que ce modèle ressemble à celui qui s'applique aux enquêtes-ménages, il y a des différences. La principale est que les répondants aux enquêtes-entreprises doivent normalement accéder à au moins une source d'information externe telle que des dossiers financiers ou administratifs.

L'aptitude des répondants à extraire les renseignements requis dépend de la mesure dans laquelle ils connaissent et comprennent la source d'information externe. Ils doivent aussi comprendre le lien entre les questions de l'enquête et la source de données externe. La multiplicité des sources de renseignements peut ajouter à la difficulté ou à la complexité de cette tâche. La complexité augmente encore si le répondant doit consulter une autre personne capable de fournir les renseignements demandés et qui peut elle-même devoir utiliser une ou plusieurs sources de données (Gower et Nargundkar 1991).

4. ELABORATION ET ESSAI DES QUESTIONNAIRES D'ENQUÊTES-ENTREPRISES

L'élaboration et l'essai des questionnaires d'enquêtes-entreprises comporte plusieurs étapes de base, dont on traite ci-dessous.

4.1 Détermination des objectifs et des besoins en données

Il faut préparer un document dans lequel les objectifs de l'enquête, les besoins en données et le plan d'analyse des données seront énoncés clairement et en détail. Ce document constitue une étape nécessaire qui mène à la détermination des variables à mesurer, des questions de l'enquête et des diverses réponses possibles. Au moment de la conception du questionnaire, il importe de déterminer et de comprendre pour quelle raison chaque question est posée, comment les renseignements recueillis seront utilisés et si les questions seront de bonnes mesures de ce qui est recherché.

4.2 Consultation avec les clients, avec les utilisateurs des données, avec les spécialistes et avec les répondants

Au moment de formuler les objectifs et les besoins en données, il faut consulter les clients et les utilisateurs des données pour bien comprendre leurs besoins et leurs attentes. Il faut aussi obtenir les conseils et l'aide de spécialistes. Dans la mesure du possible, le chercheur qui effectue une enquête doit consulter les membres de la population visée par l'enquête. Cette démarche aidera à déterminer quelles questions et préoccupations sont importantes pour les répondants et pourra influencer les décisions relatives au contenu du questionnaire. De plus, la consultation des répondants permettra de déterminer le langage et la terminologie que les répondants eux-mêmes utilisent et aidera à préciser la terminologie, les concepts et les définitions.

4.3 Questionnaires antérieurs

L'examen des questions d'autres enquêtes sur le même sujet ou sur un sujet semblable est un point de départ utile pour formuler les questions et les catégories de réponses. Dans certains cas (p. ex. pour comparer les données dans le temps), les mêmes questions peuvent être utilisées. Le chercheur doit s'assurer que les questions sont libellées de façon à permettre d'obtenir des mesures valables, cohérentes et effectives des variables étudiées.

4.4 L'utilisation de groupes de discussion pour l'élaboration de questionnaires

Un groupe de discussion traite de façon informelle d'un sujet donné et est composé de personnes choisies dans la population visée par l'enquête. Cette technique permet de recueillir des observations sur les attitudes, opinions, préoccupations et expériences des participants. Un groupe de discussion est dirigé par un animateur qui connaît bien les techniques de discussion de groupe et l'objet de la discussion.

Conception des questionnaires d'enquêtes-entreprises

A.R. GOWER¹

RÉSUMÉ

Cet article donne un aperçu de considérations importantes qui doivent entrer en ligne de compte dans l'élaboration et la conception des questionnaires d'enquêtes-entreprises. Ces considérations sont la détermination des objectifs et des besoins en données, la consultation des utilisateurs des données et des répondants et les méthodes à utiliser pour faire l'essai des questionnaires. Pour l'élaboration et de la conception d'un questionnaire d'enquête-entreprise, les groupes de discussion et les méthodes de recherche cognitive aident le chercheur à découvrir des sources possibles d'erreur de mesure et à comprendre le processus de réponse suivi par le répondant qui doit remplir le questionnaire. Cet article donne également des exemples d'utilisation de groupes de discussion et de travaux de recherche cognitive à Statistique Canada.

MOTS CLÉS: Essai de questionnaires; groupes de discussion, recherche cognitive.

1. INTRODUCTION

Il existe de nombreux types de questionnaires d'enquêtes-entreprises. Généralement, un questionnaire d'enquête-entreprise sert à recueillir des renseignements sur le personnel, les stocks, les intrants, les produits, les ventes et les finances d'une entreprise. Il peut aussi permettre de recueillir des renseignements qui ont trait à une étude de marché ou à la mesure de la satisfaction de clients. Les enquêtes-entreprises sont réalisées soit par la poste soit par un intervieweur qui pose les questions sur place ou au téléphone. Les suivis d'enquêtes postales sont souvent effectués par téléphone. Les nouvelles techniques de collecte des données d'enquêtes-entreprises sont l'interview assistée par ordinateur, les télécopieurs, l'auto-réponse au moyen d'un téléphone à clavier et la transmission électronique des données.

Comme dans les autres types d'enquêtes, la question-naire a, dans l'enquête-entreprise, un rôle important dans le processus de collecte des données. Il a une très grande influence sur la qualité des données et sur l'image qu'a aux yeux des répondants un organisme qui réalise des enquêtes. L'objet de cet article est de donner un aperçu de la conception des questionnaires d'enquêtes-entreprises. On y traite de préoccupations importantes telles que la détermination des objectifs et des besoins en données, la consultation des utilisateurs des données et des répondants, les caractéristiques et les préoccupations des répondants aux enquêtes-entreprises et les méthodes à utiliser pour faire l'essai des questionnaires.

Pour élaborer et concevoir des questionnaires d'enquêtes-entreprises, il importe particulièrement de comprendre le processus de réponse suivi par les répondants pour remplir les questionnaires. Cet article montre donc qu'il est efficace d'utiliser des groupes de discussion et des techniques de recherche cognitive pour élaborer des questionnaires d'enquêtes-entreprises et en faire l'essai. On y donne des

exemples d'utilisation de groupes de discussion et de travaux de recherche cognitive au Centre d'information sur la conception des questionnaires de Statistique Canada.

2. QUESTIONNAIRES D'ENQUÊTES-ENTREPRISES

Dans une enquête-entreprise, un questionnaire bien conçu doit permettre de recueillir des données efficaces-mment et avec un minimum d'erreurs. Il faut en outre faciliter le codage et la saisie des données, réduire au minimum le contrôle et l'imputation et permettre une réduction globale du coût et du temps associés à la collecte et au traitement des données (Statistique Canada 1994).

De nombreuses considérations entrent en ligne de compte dans l'élaboration et la conception des questionnaires d'enquêtes-entreprises. Une des plus importantes est la nature de la population des répondants. Les personnes qui répondent à des enquêtes-entreprises le font en leur qualité d'employeurs ou d'employés d'une entreprise. La façon de remplir un questionnaire dépend du poste et du niveau de responsabilités du répondant dans l'entreprise ou la société. Pour une enquête-entreprise, il est donc très important de déterminer quelle personne est le mieux placée pour fournir les renseignements.

Le fardeau de déclaration est une préoccupation très réelle pour les répondants aux enquêtes-entreprises. Il dépend du nombre de questions posées, du temps nécessaires pour remplir le questionnaire et de l'effort que les répondants doivent consacrer à faire des recherches dans d'autres sources de données ou à manipuler les données qui s'y trouvent pour fournir les renseignements dans la présentation voulue. La taille des entreprises varie. Une grande entreprise peut avoir des employés dont une des fonctions est de remplir des formulaires des gouvernements et des questionnaires

¹ A.R. Gower, Centre d'information sur la conception des questionnaires, Statistique Canada, Ottawa (Ontario) K1A 0T6.

Estimation de totaux de population

Une fois obtenue l'estimation de la taille de la base de sondage, il arrive souvent que cette estimation soit combinée à une moyenne d'échantillon pour donner une estimation d'un total de population ($\bar{Y} = N\bar{y}$). L'estimation du total de population peut comporter un biais et une variance additionnelle du fait que N est estimé. L'estimation peut aussi être biaisée parce que N n'est pas fondé sur un échantillon aléatoire de la base complète.

Plans de sondage plus complexes

Dans le présent article, nous avons examiné l'estimation de la taille de la base de sondage dans un contexte d'échantillonnage aléatoire simple, au moyen de la méthode de saisie-ressaisie. D'autres questions se posent si des plans de sondage plus complexes sont utilisés. Par exemple, dans le cas des plans stratifiés, on peut se demander s'il y a lieu d'estimer la taille de la base dans chaque strate séparément, ou d'estimer la taille de la base totale pour ensuite répartir celle-ci entre les strates en supposant des probabilités égales de présence des différentes strates dans les listes incomplètes. Une question encore plus complexe à trait à la façon d'estimer la taille de la base dans les plans de sondage à plusieurs degrés. Il est clair que ces domaines devront faire l'objet de recherches additionnelles.

REMERCIEMENTS

Les auteurs tiennent à remercier le rédacteur en chef, le rédacteur associé et deux arbitres anonymes pour leurs commentaires utiles qui ont permis d'améliorer sensiblement le contenu de cet article.

BIBLIOGRAPHIE

- CHAPMAN, D.G. (1951). Some properties of the hypergeometric distribution with application to zoological census. *University of California Publication in Statistics*, 1, 131-160.
- COCHRAN, W.G. (1978). *Sampling Techniques* (3^{ème} édition). New York: John Wiley and Sons.
- COWAN, C.D., BREAKER, W.R., et FISCHER, P.J. (1986). The methodology of counting the homeless. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-175.
- DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.
- FIEINBERG, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18, 143-154.
- GOODMAN, L.A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55, 708-713.
- GREENE, M.A. (1983). Estimating the size of the criminal population using an open population approach. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-13.
- HARTLEY, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- HARTLEY, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, C, 36, 99-118.
- JOLLY, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52, 225-247.
- KENDALL, W.L. (1992). Robust Design in Capture-Recapture Sampling: Modeling Approaches and Estimation Methods. Unpublished PhD. dissertation, North Carolina State University, Biomathematics Program.
- MENKINS, G.E., et ANDERSON, S.H. (1988). Estimation of small mammal population size. *Ecology*, 69, 1952-1959.
- OTTIS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-125.
- POLLOCK, K.H. (1974). The Assumption of Equal Catchability of Animals in Tag-Recapture Experiments. Unpublished Ph.D. dissertation, Cornell University, Biometrics Unit.
- POLLOCK, K.H. (1982). A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, 46, 752-757.
- POLLOCK, K.H., NICHOLS, J.D., HINES, J.E., et BROWNIE, C. (1990). Statistical inference for capture-recapture experiments. *Wildlife Monographs*, 107, 1-97.
- POLLOCK, K.H., et OTTO, M.C. (1983). Robust estimation of population size in closed animal populations from capture-recapture experiments. *Biometrics*, 39, 1035-1049.
- POLLOCK, K.H. (1991). Modelling, capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. *Journal of the American Statistical Association*, 86, 225-238.
- SCHNABEL, Z.E. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, 45, 348-352.
- SEBER, G.A.F. (1965). A note on the multiple recapture census. *Biometrika*, 52, 249-259.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters* (2^{ème} édition). New York: MacMillan.
- SUDMAN, S., SIRKEN, M.G., et COWAN, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-995.
- VOGEL, F.A., et KOTT, P. (1993). Multiple frame establishment surveys. *Proceedings, International Conference on Estabishment Surveys*.
- WHITE, G.C., ANDERSON, D.R., BURNHAM, K.P., et OTIS, D.L. (1982). *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos, NM: Los Alamos Laboratory.
- WITTES, J.T. (1974). Applications of a multinomial capture-recapture model to epidemiological data. *Journal of the American Statistical Association*, 69, 79-93.
- WITTES, J.T., COLTON, T., et SIDEL, V.W. (1974). Capture-recapture method for assessing the completeness of case ascertainment when using multiple information sources. *Journal of Chronic Diseases*, 27, 25-36.

dans certaines périodes, puis reprendre son activité. Le modèle de Jolly-Seber fait l'hypothèse que les bateaux de pêche qui cessent leur activité ne reviennent pas. Cette question exige un examen plus approfondi. L'utilisation du plan robuste (c.-à-d. d'une combinaison de modèles à population fermée et de modèles à population ouverte) permet de tenir compte de l'émigration temporaire. Un tel plan exigerait l'établissement de deux listes dans un court intervalle à chaque période.

3.2.2 Modèles à population fermée

Si les estimations du modèle de Jolly-Seber des "taux de survie" et du "recrutement" laissaient croire que la population est fermée (c.-à-d. que N demeure constant), les modèles généraux à population fermée présentés à la section 2.2 pourraient s'appliquer. Les avantages sont une précision accrue de N attribuable à l'utilisation d'un plus grand nombre de listes, ainsi qu'une robustesse accrue de N à l'égard des probabilités inégales de saisie. L'inconvénient principal est un accroissement de la complexité.

4. ANALYSE

4.1 Méthodes à employer quand les listes sont incomplètes

(i) Compléter la liste

L'avantage est que le chercheur dispose d'une base complète et n'a pas besoin de généraliser les résultats en fonction d'une taille estimée de la base de sondage. L'inconvénient est le coût et, parfois, l'impossibilité de compléter la liste.

(ii) Utiliser une base areolaire

Cette solution est avantageuse du fait qu'on doit seulement dénombrer les établissements des régions qui servent à former l'échantillon. Elle pourrait en revanche se révéler inefficace si les entreprises étaient dispersées à l'intérieur de chaque grande région.

(iii) Utiliser une liste et une base areolaire (méthode à bases multiples)

Les avantages sont évidemment une précision accrue et une couverture de tous les établissements. Cette méthode peut par contre se révéler coûteuse ou encore être impossible à mettre en oeuvre.

(iv) Utiliser la saisie-ressaisie pour l'estimation de la taille de la base de sondage

Cette méthode pratique a l'avantage d'être moins coûteuse que les trois précédentes. Son inconvénient est l'introduction possible d'un biais si les hypothèses de la méthode de saisie-ressaisie ne sont pas respectées et s'il faut inclure une variation attribuable à l'estimation de la taille de la base de sondage dans les estimations de la variance des estimateurs de totaux de population.

Hypothèses des modèles

(i) Population fermée

La taille de la base de sondage peut-elle être considérée comme constante, de telle sorte qu'on puisse utiliser les modèles à population fermée? La réponse variera selon que l'enquête est seulement un cliché pris à un moment précis ou une série d'observations s'étalant sur une certaine période. Elle dépendra aussi de la rapidité avec laquelle les établissements ferment leurs portes ou de la rapidité d'apparition de nouveaux établissements. Nous croyons que c'est le type d'établissement étudié qui déterminera s'il y a lieu d'utiliser des modèles à population fermée ou des modèles à population ouverte.

4.2 Estimation de la taille de la base par la méthode de saisie-ressaisie

Dans la présente section, nous examinons les hypothèses des modèles, la précision des estimations, l'estimation de totaux de population et les problèmes spéciaux propres aux plans d'échantillonnage plus complexes dans les cas où la méthode de saisie-ressaisie est appliquée à l'estimation de la taille de la base de sondage.

(iii) Perte de marques – Identification unique des établissements

Comme nous l'avons dit plus haut, les listes utilisées doivent idéalement être indépendantes, de façon que les estimations de la taille de la base de sondage ne soient pas biaisées. En pratique, il peut se révéler difficile de trouver deux listes ou plus qui soient indépendantes.

(iii) "Capitabilité inégale" et indépendance des listes

Il y a aussi la question de l'émigration temporaire, c.-à-d. de la sortie de la base de certains établissements, qui y reviennent ensuite. Il s'agissait d'un problème possible dans l'exemple des bateaux de pêche, car certains bateaux cessent leur activité, puis redeviennent actifs. Le problème pourrait se poser dans d'autres enquêtes auprès des établissements, si les entreprises arrêtent et reprennent leur activité fréquemment et gardent le même nom quand elles redeviennent actives.

Précision des estimations

Les listes utilisées doivent être de taille suffisante pour permettre une précision adéquate de l'estimation (N) de la taille de la base de sondage. Seber (1982, p. 96) examine l'estimation de Lincoln-Petersen en détail et présente des graphiques des tailles d'échantillon requises pour différents niveaux de précision. Pollock et coll. (1990) présentent des données sur la taille des échantillons dans le cas des modèles à population ouverte.

Estimations de l'effort total

L'effort total et les prises totales ont été estimés à une fréquence hebdomadaire. À titre d'illustration, nous présentons ici les calculs pour la semaine du 8 au 14 juin 1992 en ce qui a trait à l'effort total.

Effort total – bateaux privés

$N_c = 2,519$ bateaux, $\widehat{\text{Var}}(N_c) = 91,856,4706$, $\varepsilon = 0.15108$ sorties par interview, $\widehat{\text{Var}}(\varepsilon) = 0.001242$ et $\widehat{ET}(\varepsilon) = 0.0352$. En utilisant ces estimations, on obtient

$$\begin{aligned} E &= N_c \times \varepsilon = 2,519 \times 0.15108 = 380.57 \text{ sorties,} \\ \widehat{\text{Var}}(E) &= \widehat{\text{Var}}(\varepsilon)(N_c^2) + \widehat{\text{Var}}(N_c)(\varepsilon)^2 + \\ &\quad \widehat{ET}(\varepsilon) = 100.45. \end{aligned}$$

Il est utile de calculer également la variance de l'effort total, en supposant connue la taille de la base de sondage. Dans ce cas, on a $\text{Var}(E) = 7,780,9384$ avec $\widehat{ET}(E) = 88.77$; autrement dit, 89% de l'erreur-type de l'estimation de l'effort total est attribuable à la variation de l'effort moyen, tandis que seulement 11% en est attribuable à l'estimation de la taille de la base.

Effort total – bateaux affrétés

Pour les bateaux affrétés, $E = 59.95$ sorties avec $\widehat{\text{Var}}(E) = 512.5100$ et $\widehat{ET}(E) = 22.64$.

La variance de l'estimation de l'effort total, si l'on suppose connue la taille de la base de sondage, est $\text{Var}(E) = 404.8926$ avec $\widehat{ET}(E) = 20.12$. Encore ici, 89% de l'erreur-type de l'estimation de l'effort total est attribuable à la variation de l'effort moyen, tandis que seulement 11% en est attribuable à l'estimation de la taille de la base.

3.2 Utilisation de plus de deux listes

À la section 2, nous avons indiqué qu'il existe beaucoup plus de possibilités de modélisation si l'on dispose de plusieurs listes (plus de deux). Nous allons examiner ici des modèles à population fermée et à population ouverte, dans le contexte le plus général. Nous entrevoions le plan d'échantillonnage de la façon suivante. Avant le début de la saison de pêche, un échantillon préliminaire serait prélevé pour établir une liste (interviews téléphoniques ou bateaux interceptés à quai). Au cours de chaque période (par exemple de deux semaines), on établirait une liste additionnelle au moyen d'une enquête téléphonique ou d'interviews réalisées à quai. Chaque bateau aurait alors un historique de saisie qui indiquerait sur quelles listes il figure. (Par exemple, pour cinq périodes, un historique de saisie de 1 1 0 1 indiquerait que le bateau figure sur toutes les listes sauf celle de la quatrième période.) La structure de l'échantillon et de la population serait alors celle indiquée au tableau 1. La première question à examiner est de savoir si nous devons utiliser des modèles à population fermée ou à population ouverte. La façon

de procéder la plus logique consiste à ajuster le modèle à population ouverte de Jolly-Seber d'abord, puis à l'utiliser pour évaluer l'hypothèse d'une population fermée.

Tableau 1

Structure de la population selon un modèle

Période	Liste préalable					Listes établies pendant la saison (p. ex. toutes les deux semaines)				
	0	1	2	3	k	0	1	2	3	k
Tailles de la population marquée	M_0	M_1	M_2	M_3	M_k					
Tailles de la population totale	N_0	N_1	N_2	N_3	N_k					

* Les tailles de la population marquée et de la population totale sont indiquées pour l'ensemble de l'étude.

3.2.1 Modèles à population ouverte

Dans le modèle de Jolly-Seber déjà examiné à la section 2.3, un certain nombre de paramètres sont identifiables (tableau 2). Notons qu'il est possible d'estimer le nombre de bateaux de pêche de la flotte à toutes les périodes de la saison, sauf la dernière (c.-à-d. que N_k ne peut être estimé). Cette application du modèle, qui comporte l'utilisation d'une liste préalable, est avantageuse du fait qu'elle écarte toute inquiétude quant à la possibilité que la liste préalable soit périmée, car le modèle permet les ajouts et les suppressions avant le début de la saison. Le modèle de Jolly-Seber est toutefois désavantageux en raison de sa complexité plus grande. Pour chaque période, une taille de la base de sondage est établie et les paramètres du taux de survie et du recrutement doivent être estimés. Parfois, les estimations de ces paramètres ont une faible précision, à moins que les échantillons ne soient de grande taille. Le modèle de Jolly-Seber a aussi l'inconvénient d'exiger l'hypothèse d'une probabilité égale de saisie.

Tableau 2

Structure du modèle à population ouverte de Jolly-Seber*

Période	Valeurs préalables					Valeurs établies pendant la saison				
	0	1	2	3	k-1	k				
Population marquée ($M_0 = 0$)	M_1	M_2	M_3			M_{k-1}				
Population totale	-	N_1	N_2	N_3		N_{k-1}				
Taux de survie	\hat{p}_0	\hat{p}_1	\hat{p}_2			\hat{p}_{k-2}				
Recrutement		B_1	B_2			B_{k-2}				

* Les estimateurs des paramètres identifiables sont indiqués pour la taille de la population marquée, la taille de la population totale, le taux de survie et le recrutement.

Une autre question importante qui se pose au sujet du modèle de Jolly-Seber a trait à ce qu'on appelle l'"émigration temporaire". Un bateau de pêche peut devenir inactif

marquées la probabilité de capture est plus élevée ou plus faible que pour les individus non marqués. Dans les deux cas, lorsque les unités des listes sont des bateaux de pêche, nous croyons qu'il peut exister une hétérogénéité des probabilités de saisie entre les bateaux. Si l'hétérogénéité vaut pour les deux échantillons, les individus saisis sur la première liste auront tendance à être ceux pour lesquels la probabilité de saisie est élevée et donc ceux qui le plus probablement figureront sur la deuxième liste. La proportion d'individus marqués dans le deuxième échantillon (liste) sera donc trop élevée et l'estimateur de N comportera un biais négatif. Notons que cet argument intuitif montre clairement que ce n'est pas l'hétérogénéité en soi qui constitue un problème, mais la corrélation positive des probabilités de saisie entre les deux échantillons. Une autre façon d'énoncer l'hypothèse des probabilités égales de saisie consiste à dire que les probabilités de saisie sont indépendantes entre les deux échantillons. Une méthode par laquelle on peut tenter de réaliser l'indépendance des probabilités de saisie entre les deux échantillons consiste à utiliser des plans de sondage complètement différents pour les deux échantillons. C'est pourquoi nous avons recommandé plus tôt qu'une liste soit établie à partir d'interviews téléphoniques et l'autre, à partir d'interviews réalisées à quai. Toutefois, nous croyons qu'il reste une hétérogénéité et une dépendance dans les probabilités de saisie: nous pensons que pour les bateaux qui participent très activement à la pêche la probabilité de figurer sur l'une ou l'autre liste est plus grande (liste d'interviews téléphoniques et liste de bateaux à quai). Cette hétérogénéité causera un biais négatif dans l'estimation de la taille de la base de sondage, mais nous n'avons aucune idée de l'importance de ce biais négatif. Seber (1982, p. 86) fait un examen plus complet de l'hétérogénéité et de l'indépendance des échantillons.

Perte ou non-détection des marques

La situation est ici un peu confuse. On pourrait croire, au premier abord, que la perte ou la non-détection de marques est impossible dans cette application. Cela suppose toutefois que les bateaux ont tous des noms distincts ou que, si certains ont des noms identiques, d'autres renseignements comme le nom du capitaine permettent de distinguer toutes les unités figurant sur les listes. S'il n'est pas certain que l'on puisse identifier tous les bateaux, il n'est pas sûr non plus qu'on puisse confirmer si un bateau marqué a été ressaisi ou non. Il se peut aussi que les agents introduisent des erreurs, rendant ainsi difficile l'appariement d'une ressaisie avec son identification dans la liste initiale. Un mode d'emploi normalisé est actuellement mis au point et documenté, en vue de réduire au minimum les erreurs de ce genre dans l'avenir.

3.1.2 Estimation de l'effort total et des prises totales
L'effort total (E) (c.-à-d. le nombre total de sorties de pêche au cours d'une période définie) est ainsi estimé:
(3.1) $E = Ne$

3.1.3 Illustration de la méthode

Les prises totales (C) sont estimées par $C = Ec$, où E est l'effort de pêche total estimé et c , les prises moyennes par unité d'effort calculées d'après les interviews à quai. Les mêmes préoccupations que celles touchant l'équation (3.1) s'appliquent sans doute à cette équation et, encore une fois, une simulation pourrait être très utile.

$$\widehat{\text{Var}}(E) = (N)^2 \widehat{\text{Var}}(e) + (e)^2 \widehat{\text{Var}}(N) + \widehat{\text{Var}}(e) \widehat{\text{Var}}(N). \quad (3.2)$$

ainsi la variance estimée de E :
où N est l'estimation de la taille de la base (taille de la flotte) et e , l'effort de pêche moyen (c.-à-d. le nombre moyen de sorties de pêche) obtenu au moyen des interviews téléphoniques. L'évaluation des propriétés de cet estimateur est plus difficile que quand N est connu, car N et e sont deux variables aléatoires. Nous croyons que e est biaisé vers le haut, car les bateaux de pêche peu actifs sont moins susceptibles de figurer sur la liste. Malheureusement, nous ne pouvons dire que N sera toujours biaisé vers le haut ou vers le bas. Les trois violations des hypothèses examinées en 3.1.1 peuvent être importantes (population fermée, hétérogénéité, perte de marques), et l'on ne sait trop dans quel sens le biais global de N se manifesterait. La seule méthode possible est de faire une simulation avec plusieurs scénarios différents quant aux violations des hypothèses. En utilisant l'équation (1.4), on peut exprimer

Dans cette section, nous présentons les estimations de la taille de la base de sondage et les estimations de l'effort total pour la pêche au thon rouge en Virginie pendant une partie de l'année 1992. Ces estimations proviennent d'une enquête plus vaste couvrant la côte est des Etats-Unis, de la Caroline du Nord au Massachusetts. Les estimations sont distinctes pour les bateaux affrétés et les bateaux privés.

Estimations de la taille de la base de sondage

Des listes de bateaux privés et de bateaux affrétés ont été établies, principalement d'après les interviews téléphoniques des saisons précédentes. Au cours de la saison 1992, des bateaux "marqués" et "non marqués" ont été échantillonnés aux pompes à carburant avant ou après les sorties de pêche.
Pour les bateaux privés, la taille de la liste était de $M = 335$ bateaux avant la saison. Un échantillon de $n = 374$ bateaux ont été interceptés aux pompes à carburant et, de ces derniers, $m = 49$ étaient marqués. L'estimateur de Chapman est $N_c = 2,519$, $ET(N_c) = 303.08$ et ET relative = 0.12.
Pour les bateaux affrétés, la taille de la liste était de $M = 47$ bateaux avant la saison. Un échantillon de $n = 31$ bateaux ont été interceptés aux pompes à carburant et, de ces derniers, $m = 13$ étaient marqués. L'estimateur de Chapman est $N_c = 109$, avec $ET(N_c) = 17.88$ et ET relative = 0.16.

2.4 Combinaison de modèles à population fermée et à population ouverte

Pollock (1982), Pollock et coll. (1990) et Kendall (1992) examinent des méthodes d'échantillonnage qui permettent de réunir dans un même plan des modèles à population fermée et des modèles à population ouverte. Ces méthodes présentent un avantage du fait qu'on peut considérer que les probabilités de capture sont inégales, possibilité que n'offre pas le modèle classique de Jolly-Seber. Autre avantage, ces modèles permettent de tenir compte d'une émigration temporaire des animaux.

2.5 Applications des modèles de saisie-ressaisie

Naturellement, les modèles de saisie-ressaisie ont beaucoup été utilisés pour l'étude de la faune terrestre et des poissons. Toutefois, on peut maintenant observer plusieurs applications nouvelles, en dehors du domaine biologique. De nombreux auteurs ont appliqué les méthodes de saisie-ressaisie à l'estimation du sous-dénombrement du recensement. (Voir Feinberg (1992) pour une bibliographie complète.) Cowan, Breakay et Fischer (1986) s'en sont servi pour estimer le nombre de sans-abri dans une ville. Greene (1983) a utilisé la méthode pour estimer des paramètres démographiques propres à des populations criminelles. Wittes (1974) et Wittes, Colton et Sidel (1974) ont utilisé la saisie-ressaisie pour estimer le nombre de personnes atteintes de maladies à partir de listes d'hôpital et d'autres listes. L'échantillonnage de populations humaines difficiles à circonscrire, effectué à l'aide de l'échantillonnage par grappes, d'échantillons superposés et de l'échantillonnage par saisie-ressaisie, a été examiné par Sudman, Sirken et Cowan (1988).

3. MODÈLES DE SAISIE-RESSAISIE APPLIQUÉS À LA "LARGE PELAGIC SURVEY"

La "Large Pelagic Survey" est une enquête sur la pêche sportive réalisée par le National Marine Fisheries Service au moyen d'un plan d'enquête combinant une enquête téléphonique et une enquête aux points d'accès. Un échantillon de propriétaires de bateaux de pêche est constitué à partir d'une liste et l'on demande à ces propriétaires, par téléphone, de fournir des renseignements sur leur effort de pêche (nombre de sorties de pêche au cours d'une période donnée). Les données sur les prises par unité d'effort (prises par sortie) sont tirées d'un deuxième échantillon de propriétaires de bateaux interviewés aux points d'accès des bateaux, au retour des sorties de pêche. On combine les données des deux enquêtes pour estimer l'effort total et la prise totale d'espèces importantes comme le thon rouge.

Le fait que la liste des propriétaires de bateaux qui sert à l'enquête téléphonique soit très incomplète pose un problème sérieux dans cette enquête. Par conséquent, la théorie classique de l'échantillonnage, qui suppose une base complète de taille connue (N), ne convient pas et doit être modifiée. La méthode actuelle d'estimation de la taille

Captabilité égale

Le non-respect de l'hypothèse de la captabilité égale des individus peut être dû soit à l'hétérogénéité intrinsèque des probabilités de capture entre les individus, soit à une "réaction à la capture" qui ferait que pour les individus

Population fermée

Selon la méthode actuelle, les bateaux "marqués" (M) sont ceux de la liste principale, établie en majeure partie d'après des interviews téléphoniques antérieures. L'échantillon de ressaisie est prélevé à quai, aux pompes à carburant, et l'on vérifie la liste complète des bateaux interceptés (n) afin de déterminer ceux qui sont "marqués" (m), (c.-à-d. ceux qui appartiennent à la liste principale). On obtient ensuite un estimateur de la taille de la base de sondage (N) en utilisant l'équation 2.3. Examinons maintenant les hypothèses du modèle et l'effet que la violation de ces dernières pourrait avoir sur le biais de l'estimateur de N .

3.1 Le modèle de Lincoln-Petersen

de la base de sondage sous-jacente à la liste consiste à combiner deux listes (une liste téléphonique et une liste de bateaux interceptés à quai) et à utiliser le modèle de Lincoln-Petersen. On peut se demander si cette façon de procéder est la meilleure. Il serait peut-être possible, par exemple, de combiner plus de deux listes et ensuite d'utiliser les modèles à population fermée ou ouverte dont il a été question aux sections 2.2 et 2.3. Toutefois, nous laissons de côté ces questions pour le moment et nous commençons par examiner et évaluer la méthode actuelle, afin de montrer comment cette approche pourrait être employée utilement dans d'autres enquêtes auprès des établissements.

3.1.1 Estimation de la taille de la base de sondage (N)

Cette hypothèse n'est vraisemblablement pas respectée. Il se peut que des bateaux de pêche aient été inscrits sur la liste principale puis aient cessé leur activité de pêche (pertes), tandis que de nouveaux bateaux peuvent être entrés en activité une fois la période commencent (gains). Idéalement, il faudrait produire une estimation distincte de la taille de la base de sondage pour chaque période de deux semaines. L'avantage de l'estimateur fondé sur un modèle à population fermée de Lincoln-Petersen est qu'il est simple et pratique à utiliser. Le fait que la population ne soit pas fermée peut entraîner un biais positif ou négatif de l'estimateur.

On ne connaît pas pour le moment l'évolution probable de la taille de la flotte pendant la saison de pêche. Un plan de sondage à saisies-ressaisies multiples permettrait d'utiliser le modèle de Jolly-Seber pour estimer la taille de la flotte à chaque période. L'examen de ces estimateurs, ainsi que des estimateurs du taux de survie et du recrutement, nous permettra d'évaluer la validité de l'hypothèse d'une population fermée. Pour l'instant, nous devons nous limiter à des conjectures.

L'expression suivante donne une estimation de la variance de N_c :

$$\widehat{\text{Var}}(N_c) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)} \quad (2.4)$$

Voir par exemple Seber (1982, p. 60).

Les hypothèses fondamentales de ce modèle sont les suivantes:

- (a) la population est complètement fermée aux ajouts et aux suppressions;
- (b) tous les poissons ont une probabilité égale d'être capturés dans chaque échantillon;
- (c) les marques restent en place et sont détectées.

L'hypothèse relative à la population fermée peut être adoucie, mais même pour une population complètement ouverte, à laquelle cet estimateur ne s'applique pas, une modification de l'estimateur de Lincoln-Petersen est utilisée. L'hypothèse d'une capturabilité égale pose des problèmes dans la plupart des applications. Il pourrait bien y avoir une variabilité inhérente (hétérogénéité) dans les probabilités de capture des animaux individuels en raison du sexe ou d'autres facteurs. Il se peut aussi qu'une réaction à la capture initiale se manifeste. Dans la section qui suit, nous examinons des modèles à population fermée comportant plus de deux échantillons, dans lesquels les probabilités de capture des animaux peuvent être l'objet d'une variation dans le temps ainsi que d'une hétérogénéité et de réactions à la capture. La perte et la non-détection des marques peuvent être importantes. Une façon d'estimer la perte des marques consiste à utiliser deux marques (Seber 1982, p. 94).

2.2 Modèles à population fermée

Les modèles à population fermée reposent sur l'hypothèse qu'aucun mouvement de la population attribuable à des naissances, à des décès ou à une migration ne survient entre les périodes d'échantillonnage. Ces modèles, par conséquent, sont généralement appliqués à des études portant sur de courtes périodes (p. ex. on pose des filets chaque jour pendant cinq jours consécutifs). Les historiques de capture des animaux individuels sont les données qui servent à produire les estimations selon ces modèles. Parmi les premiers travaux dans ce domaine, mentionnons ceux de Schnabel (1938) et de Darroch (1958), qui ont examiné des modèles fondés sur une capturabilité égale des animaux dans chaque échantillon.

On possède aujourd'hui un ensemble de modèles dans lesquels les probabilités de capture diffèrent à cause de l'hétérogénéité (h), d'une réaction à la capture (b), d'une variation dans le temps (t) (c.-à-d. que la probabilité de capture au temps i diffère de la probabilité de capture au temps j) et de toutes les combinaisons de deux ou de trois de ces facteurs. Les huit modèles $[M(o), M(h), M(b), M(bh), M(t), M(th), M(tb), M(thb)]$ ont d'abord été

- (a) pour tous les animaux présents dans la population au moment d'un échantillonnage particulier, la probabilité d'être capturés est la même;
- (b) pour tous les animaux marqués qui sont présents dans la population immédiatement après un échantillonnage, la probabilité de survivre jusqu'à l'échantillonnage suivant est la même;
- (c) les marques restent en place et sont détectées;
- (d) toute l'émigration est permanente;
- (e) tous les échantillons sont instantanés, et chaque animal est relâché immédiatement après avoir été prélevé.

Selon le modèle de base de Lincoln-Petersen, décrit à la section 2.1, les hypothèses (a), (c) et (e) devaient être faites. Seuls les animaux marqués servent à estimer les taux de survie; il n'est donc pas nécessaire, à strictement parler, de supposer l'égalité des taux de survie des animaux marqués et des animaux non marqués. En pratique toutefois, le biologiste voudra que les estimations des taux de survie portent sur l'ensemble de la population. Le modèle de Jolly-Seber permet que certains animaux soient perdus au moment de la capture et ne soient donc pas retournés dans la population. Le modèle exige par ailleurs que toute émigration soit permanente. Si des animaux émigrent, puis reviennent dans la population, il en résulte ce qu'on appelle une émigration temporaire, qui contredit gravement les hypothèses et engendre un biais important dans les estimations de la taille de la population.

2.3 Modèles à population ouverte

Dans beaucoup d'études de saisie-ressaisie, on ne peut supposer que la population est fermée, c.-à-d. à l'abri d'ajouts ou de suppressions permanentes. Le modèle de base à population ouverte qui convient à cette situation est celui de Jolly-Seber (Jolly 1965; Seber 1965; Seber 1982, p. 196). Ce modèle permet l'estimation de la taille de la population à chaque échantillonnage, ainsi que l'estimation des taux de survie et du nombre de naissances entre les échantillonnages. La migration ne peut être séparée des naissances et des décès sans information additionnelle.

Le modèle de Jolly-Seber repose sur les hypothèses suivantes:

- (a) la population est complètement fermée aux ajouts et aux suppressions;
- (b) tous les poissons ont une probabilité égale d'être capturés dans chaque échantillon;
- (c) les marques restent en place et sont détectées.

L'expression suivante donne une estimation de la variance de N_c :

effectuées auprès des établissements. Dans la dernière section, nous faisons le bilan des forces et des faiblesses de la méthode de saisie-ressaisie pour l'estimation de la taille de la base de sondage dans les enquêtes auprès des établissements. Plusieurs des idées présentées ici exigeront d'être étayées par de futures recherches.

2. BREF SURVOL DES MODELES DE SAISIE-RESSAISIE

Il va de soi qu'une couverture complète des publications relatives à la saisie-ressaisie se situe hors du cadre du présent article. Pour un compte rendu plus détaillé, nous recommandons les travaux de Seber (1982), White et coll. (1982), Pollock et coll. (1990) et Pollock (1991). L'aperçu général donné par Pollock (1991) est une bonne introduction à la littérature du domaine; nous en suivons ici très étroitement la démarche. Les autres sources sont des ouvrages et des monographies qui s'adressent au lecteur qui veut pousser plus loin sa recherche et à plus de temps à sa disposition.

Nous allons examiner brièvement le modèle de Lincoln-Petersen pour deux échantillons, puis des modèles plus généraux à population fermée et à population ouverte pour plus de deux échantillons et, enfin, une méthode combinant des modèles à population fermée et à population ouverte dans un même plan d'échantillonnage. Pollock et coll. (1990, p. 9) présentent un organigramme donnant une vue d'ensemble des modèles et des liens qui existent entre eux.

2.1 Le modèle de Lincoln-Petersen

Il s'agit du modèle de saisie-ressaisie le plus ancien, le plus simple et le mieux connu. Le premier à l'utiliser fut Laplace, pour l'estimation de la population de la France. Petersen, au tournant du siècle, a été le premier à appliquer ce modèle au domaine de la pêche. Seber (1982, chapitre 3) en donne les détails dans une excellente analyse. Dans son application originale au domaine de la pêche, la méthode peut être décrite de la façon suivante. Un échantillon de M poissons est capturé, et les poissons sont marqués et relâchés. Plus tard, un deuxième échantillon de n poissons est capturé, dont m sont marqués. On obtient de façon intuitive un estimateur en supposant égaux les proportions de poissons marqués dans l'échantillon et dans la population,

$$m/n = M/N, \tag{2.1}$$

ce qui donne

$$\hat{N} = Mn/m. \tag{2.2}$$

Un estimateur modifié, atténuant le biais dans le cas des petits échantillons, est donné par Chapman (1951):

$$\hat{N}_c = [(M + 1)(n + 1)/(m + 1)] - 1. \tag{2.3}$$

fondé sur un modèle, qui repose sur l'hypothèse clé de l'indépendance des deux listes. Dès qu'on dispose d'une estimation de la taille de la population, on peut estimer un total de population pour une caractéristique quelconque si cette caractéristique est mesurée auprès d'un échantillon de la population.

$$Y = N\hat{y}, \tag{1.1}$$

où N est connu et \hat{y} est la moyenne de l'échantillon; voir par exemple Cochran (1978, p. 21). La variance de Y est donnée par

$$\text{Var}(Y) = N^2 \text{Var}(\hat{y}), \tag{1.2}$$

où

$$\text{Var}(\hat{y}) = S^2 \frac{n}{N - n},$$

S^2 est la variance de la population et $(N - n/N)$ est appelé facteur de correction pour population finie. L'estimateur (1.1) est également un estimateur sans biais du total de population.

Dans le cas qui nous occupe, l'estimateur est

$$Y = N\hat{y}, \tag{1.3}$$

où N est obtenu par la méthode de saisie-ressaisie.

Il est clair que les propriétés de l'estimateur (1.3) sont plus difficiles à évaluer, car N et \hat{y} sont deux variables aléatoires; ce n'est pas le cas de l'estimateur (1.1), où N est une valeur connue. Dans notre cas, on a une variance estimée de Y donnée par

$$\widehat{\text{Var}}(Y) = (N)^2 \widehat{\text{Var}}(\hat{y}) + (Y)^2 \widehat{\text{Var}}(N) + \widehat{\text{Var}}(Y) \widehat{\text{Var}}(N), \tag{1.4}$$

en supposant que y et N sont indépendants et en utilisant un résultat exact dû à Goodman (1960). L'estimateur (1.3) est un estimateur sans biais seulement si N et y sont des estimateurs sans biais respectivement de la taille et de la moyenne de la population, ce qui n'est habituellement pas le cas en pratique. Nous examinons l'estimateur (1.3) dans l'exemple de l'enquête sur la grande pêche pélagique présentée à la section 3.

Le reste du présent article est structuré comme suit. À la section 2, nous passons en revue la littérature relative à la saisie-ressaisie, pour donner un aperçu des types de modèles disponibles. À la section 3, nous présentons un exemple d'enquête par sondage portant sur les bateaux de pêche. (Nous considérons un bateau comme un établissement commercial.) Bien que cet exemple comporte certaines caractéristiques uniques, nous croyons qu'il est représentatif, sous plusieurs aspects, d'autres enquêtes

Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète

K.H. POLLOCK, S.C. TURNER et C.A. BROWN¹

RÉSUMÉ

Nous présentons une solution formelle, fondée sur un modèle, au problème de l'estimation de la taille de bases de sondage (listes), au moyen d'un échantillonnage avec saisie-ressaisie; cette solution a été largement utilisée pour estimer des populations d'animaux et pour le redressement des chiffres du recensement des États-Unis. Lorsqu'on dispose de deux listes incomplètes, il est facile d'estimer la taille totale de la base, en utilisant l'estimateur de Lincoln-Petersen. Il s'agit d'un estimateur fondé sur un modèle, qui repose sur l'hypothèse clé que les deux listes sont indépendantes. Dès qu'on dispose d'une estimation de la taille de la population (base de sondage), on peut estimer un total de population pour une caractéristique quelconque si cette caractéristique est mesurée auprès d'un échantillon de la population. Une analyse des propriétés de l'estimateur de Lincoln-Petersen est présentée. Nous examinons un exemple dans lequel les établissements sont des bateaux de pêche en activité au large de la côte Atlantique des États-Unis. L'estimation de la taille de la base de sondage, puis de totaux de population, à l'aide d'un modèle de saisie-ressaisie devrait avoir de vastes applications dans les enquêtes auprès des établissements, vu son côté pratique et les économies qu'elle procure. Toutefois, la possibilité que des biais soient introduits par suite du non-respect de certaines hypothèses doit être prise en considération.

MOTS CLÉS: Bases incomplètes; échantillonnage par saisie-ressaisie; enquêtes sur la pêche sportive; enquêtes téléphoniques; enquêtes aux points d'accès.

1. INTRODUCTION

Dans la théorie classique de l'échantillonnage, on suppose qu'il existe une base de sondage complète, c'est-à-dire qu'il existe, au moins théoriquement, une liste complète des unités de la population. On peut alors extraire un échantillon probabiliste de la population. Les estimateurs de paramètres de la population, par exemple une moyenne ou un total, ont alors des propriétés connues et peuvent être facilement étudiés théoriquement et numériquement. Les ouvrages sur la théorie de l'échantillonnage, comme celui de Cochran (1978), étudient cette situation et énoncent les propriétés des estimateurs pour des plans de sondage courants comme l'échantillonnage aléatoire simple, l'échantillonnage aléatoire stratifié et l'échantillonnage à plusieurs degrés (par grappes). En pratique, dans les enquêtes auprès des établissements ou des entreprises, il se peut qu'on ne dispose pas d'une base complète. Ainsi, les listes d'établissements dont disposent les associations professionnelles ou les organismes gouvernementaux sont souvent incomplètes. Une façon d'aborder ce problème consiste à utiliser la méthode à bases multiples initialement élaborée par Hartley (1962, 1974). C'est la méthode employée, notamment, dans les enquêtes sur les établissements agricoles effectuées par le National Agricultural Statistics Service (USDA) (Vogel et Koit 1993). Ces enquêtes utilisent une liste incomplète d'établissements agricoles ainsi qu'une base aréolaire dans laquelle sont dénombrés tous les établissements appartenant

Nous présentons plus loin une solution formelle à ce problème; il s'agit d'une solution fondée sur un modèle, qui fait appel à l'échantillonnage avec saisie-ressaisie. Les modèles d'échantillonnage avec saisie-ressaisie sont largement utilisés dans l'échantillonnage de populations animales (Seber 1982) et pour redresser les chiffres du recensement des États-Unis afin d'éliminer le sous-dénombrement (Feinberg 1992). Dans le cas le plus simple, celui où l'on possède deux listes incomplètes, nous considérons les unités "marquées" comme celles qui appartiennent aux deux listes et les unités non marquées comme celles qui n'appartiennent pas aux deux listes. Il est facile d'estimer la taille totale de la base au moyen de l'estimateur de Lincoln-Petersen (Seber 1982, p. 59). Il s'agit d'un estimateur

à une unité d'échantillonnage. La liste, donc, est incomplète, tandis que la base aréolaire est théoriquement complète. (On dispose d'une liste de toutes les unités aréolaires et, dans chacune de ces unités, tous les établissements agricoles peuvent théoriquement être dénombrés.) Il y a toutefois des situations où le recours à une base aréolaire peut, pour des raisons pratiques, n'être pas possible. Le chercheur peut n'avoir à sa disposition que des listes incomplètes d'établissements. En général, dans ce cas, on fusionne toutes les listes incomplètes et l'on ne tient pas compte du fait que l'ensemble puisse demeurer incomplet. Selon l'importance des carences de la liste globale, les estimations de la taille de la population et des totaux de population peuvent alors comporter un biais négatif prononcé.

¹ K.H. Pollock, North Carolina State University, Raleigh, NC 27695; S.C. Turner et C.A. Brown, National Marine Fisheries Service, Miami, FL 33149, U.S.A.

- GOWARD, S.N., MARKHAM, B., DYE, D.C., DULANEY, W., et YANG, J. (1991). Normalized difference vegetation index measurements from the advanced very high resolution radio-meter. *Remote Sensing of the Environment*, 35, 257-277.
- HAY, A.M. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9, 1395-1398.
- HOUSTON, A.G., et HALL, F.G. (1984). Use of satellite data in agricultural surveys. *Communications in Statistics Theory and Methods*, 13, 23, 2857-2880.
- VAN LANEN, H.A.J., VAN DIEPEN, C.A., REINDS, G.J., DE KONING, G.H.J., BULENS, J.D., et BREGT, A.K. (1992). Physical land evaluation methods and GIS to explore the crop growth potential and its effects within the European Communities. *Agricultural Systems*, 39, 307-328.
- MEYER-ROUX, J. (1990). Présentation du projet pilote de télé-détection appliquée aux statistiques agricoles. Conférence on the Application of Remote Sensing to Agricultural Statistics. Office for Publications of the E.C. Luxembourg.
- NEALON, J.P. (1984). Review of the multiple and area frame estimators. U.S. Department of Agriculture, Statistical Reporting Service, Report 80, Washington, D.C.
- SHARMAN, M., et de BOISSEZON, H. (1992). Action IV: de l'image aux statistiques, bilan opérationnel après deux années d'estimations rapides des superficies et des rendements potentiels au niveau Européen. Conférence on the Application of Remote Sensing to Agricultural Statistics, Belgrate. Office for Publications of the E.C. Luxembourg.
- VAN DIEPEN, C.A., WOLF, J., VAN KEULEN, H., et RAPPOLDT, C. (1989). WOFOST: A simulation model for crop production. *Soil Use and Management*, 5, 16-24.
- EUROSTAT 1991. Working party, Crop Products Statistics. Methodological reports. Document AGRI/PE/333, Luxembourg.
- FUENTES, M., et GALLEGO, F.J. (1994). Stratification and cluster estimator on an area frame by squared segments with an aligned sample. Conference on Applied Statistics to Agriculture, Kansas State University of Manhattan, KS.
- GALLEGO, F.J. (1992). Flächenschätzungen für einjährige Feldfrüchte mit Hilfe Fernerkundung. *Neue Wege raumbezogener Statistik. Forum der Bundesstatistik*, 20, 109-120. Wiesbaden: Statistisches Bundesamt.
- GALLEGO, F.J. (1994). Using a confusion matrix for area estimation with remote sensing. *Atti Convegno AIT*, Roma, 99-102.
- GALLEGO, F.J., et DELINCÉ, J. (1994). Area estimation by segment sampling. Dans *Euro-Courses Remote sensing applied to Agricultural Statistics*.
- GALLEGO, F.J., DELINCÉ, J., et RUEDA, C. (1993). Crop area estimates through remote sensing: Stability of the regression correction. *International Journal of Remote Sensing*, 14, 18, 3433-3445.
- GIOVACCINI, A. (1992). Agricultural statistics by remote sensing in Italy: an ultimate cost analysis. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgrate. Office for Publication of the E.C. Luxembourg.
- GONZÁLEZ, F., LOPEZ, S., et CUEVAS, J.M. (1991). Comparing two methodologies for crop area estimation in Spain using landsat TM images and ground gathered data. *Remote Sensing of the Environment*, 32, 29-36.

Dans le tableau 4, on compare les résultats de l'enquête par segments (observations directes sur le terrain), ceux du sondage agricole (exploitations échantillonnées par points) et les statistiques officielles pour les principales cultures du pays. On obtient les statistiques officielles en additionnant les chiffres déclarés par toutes les exploitations ou coopératives d'Etat. Il y a une assez faible différence dans les estimations de superficie pour le blé, le maïs et les pommes de terre. Il ne faut pas exclure des réponses fournies par les exploitants agricoles un biais qui s'explique par l'autococonsommation de produits agricoles.

Tableau 4

Résultats de l'enquête par segments et du sondage agricole en République tchèque (1992)

Enquête par segments	Sondage agricole		BST					
	Super- CV	Prod. CV		Super- CV	Prod. CV			
En milliers d'hectares	Super- CV	Prod. CV	Super- CV	Prod. CV	Super- CV	Prod. CV		
Blé	824	5,4	757	3,7	3,412	4,9	780	3,413
Orge	655	5,1	630	3,8	2,521	4,3	640	2,512
Colza	140	11,6	137	6,8	310	7,5	136	296
Betterave à sucre	119	11,5	127	8,1	4,172	11,0	125	3,874
Maïs	361	7,5	326	4,8	8,884	4,3	361	8,904
Pommes de terre	109	13,6	92	7,9	1,706	8,7	111	1,969

BST: Bureau statistique tchèque.

Les coefficients de variation (CV) des estimations de la superficie sont plus faibles pour le sondage agricole que pour l'enquête par segments. Cela n'est pas surprenant puisque le sondage agricole donne des renseignements à propos des champs qui se trouvent à l'extérieur des segments. Les 458 exploitations sélectionnées représentent plus de 15% de la SAU du pays. Les CV sont un peu plus élevés pour les estimations de la production que pour les estimations de la superficie (plus faibles dans le cas du maïs). Cela semble montrer que la variabilité des rendements contribue moins que la variabilité des superficies à la variabilité de la production.

6. CONCLUSIONS ET RECOMMANDATIONS

Les bases aréolaires fondées sur des grilles carrées sont une solution de remplacement pragmatique des bases aréolaires fondées sur des éléments au sol délimités par des caractéristiques physiques. Elles sont beaucoup moins chères à constituer et ne semblent pas comporter d'inconvénients importants quant aux résultats définitifs. Toutefois, un certain travail théorique reste à faire pour qu'on puisse déterminer dans quelles conditions les erreurs d'emplacement attribuables aux limites autres que physiques ont un effet négligeable sur les estimations.

L'échantillonnage de points dans des segments aréolaires est une méthode réaliste qui permet de constituer des bases de sondage pour l'échantillonnage d'exploitations

Nous remercions les diverses administrations nationales et régionales qui nous ont apporté leur collaboration pour ce travail. Nous remercions également A. Burdill et O. O'Hanlon, qui ont bien voulu réviser cet article. Les nombreuses observations faites par les arbitres ont beaucoup contribué à rendre cet article plus utile pour les lecteurs.

BIBLIOGRAPHIE

ALLEN, J.D. (1990). A look at the remote sensing applications program of the national agricultural statistics service. *Journal of Official Statistics*, 6, 4, 393-409.

AMBRÓSIO, L., ALONSO, R., et VILLA, A. (1993). Estimación de superficies cultivadas por muestreo de áreas y teledelección. *Estadística Española*, 35, 91-103.

CARFAGNA, E., RAGNI, P., ROSSI, L., et TERPESSI, C. (1991). Area frame: un Nuovo Istrumento per la Realizzazione delle Statistiche Agricole in Italia. *Contributi alla Statistica Spaziale*. University of Parma.

CARFAGNA, E., et DELINCE, J. (1992). Farm survey based on area frame sampling. The case of Emilia Romagna in 1990. Conference on the Application of Remote Sensing to Agricultural Statistics, Belgirate. Office for Publication of the E.C. Luxembourg.

CARFAGNA, E., et GALLEGO, F.J. (1994). Extrapolating intra-cluster correlation to optimize the size of segments in an area frame. Conference on Applied Statistics to Agriculture, Kansas State University, Manhattan, KS.

COCHRAN, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons.

COTTER, J., et NEALON, J. (1987). Area frame design for agricultural surveys. U.S. Department of Agriculture. National Agricultural Statistics Service.

DELINCE, J. (1990). Un premier bilan de l'action I Inventaires Régionaux du Projet Agriculture après deux années d'activité. Conference on the Application of Remote Sensing to Agricultural Statistics, Varese. Office for Publication of the E.C. Luxembourg.

DICORATO, F. (1993). AIS estimation programs. User documentation. JRC Ispra.

Le nombre d'animaux d'élevage est très sous-estimé (tableau 3) puisque beaucoup de propriétaires d'animaux d'élevage ne possèdent pas de terres agricoles. Nous avons utilisé une méthode mixte pour les bestiaux et les porcs: recensement à l'aide d'une liste, pour les cinquante plus grosses exploitations, échantillonnage par points pour les autres exploitations. La procédure est utilisable pour les porcs, mais les CV ne sont pas encore satisfaisants.

Tableau 3

Résultats du sondage agricole sur la base aréolaire et sur la base combinée pour les animaux d'élevage en Emilie-Romagne (1990)					
En milliers d'unités	Recensement	Base aréolaire		Base combinée	
		Estimation	CV %	Estimation	CV %
Bestiaux	869	829	14	894	13
Porcs	1,876	1,312	37	1,818	27
Moutons	90	38	74		

5.2 République tchèque 1992

Les bases areolaires semblent particulièrement utiles dans les anciens pays communistes d'Europe à cause du changement rapide dans la structure de la propriété foncière. Les statistiques agricoles sont le plus souvent produites sans erreur d'échantillonnage en additionnant les données déclarées par chaque exploitation ou coopérative d'Etat. Cette procédure ne pourra plus s'appliquer au cours des prochaines années. Il sera extrêmement difficile d'avoir une idée du nombre d'exploitations existantes, et un recensement de l'agriculture sera dépassé avant que les données puissent être produites. Il se peut que les bases areolaires soient la meilleure solution.

Le territoire de la République tchèque (environ 80,000 km²) a été distribué en six strates par interprétation photographique d'images obtenues par le capteur TM des satellites Landsat. Il a fallu quinze jours à une personne pour effectuer cette stratification. En 1992, on a effectué une enquête avec un échantillon de 417 segments carrés de 400 ha en répétant une structure fixe de blocs de 40 km sur 40 km. On a visité les segments et obtenu l'estimation de la superficie selon la procédure expliquée à la sous-section 2.4.1.

Les exploitations ont été échantillonnées à l'aide d'une grille fixe de cinq points dans chaque segment. La grille de cinq points avait la forme d'un "x", comme dans la figure 6. Cette procédure a permis d'obtenir 2,085 points: 858 correspondaient à des régions non agricoles tandis que les 1,227 autres provenaient de 458 exploitations. Aucune donnée manquante n'a été enregistrée: toutes les exploitations ont été repérées et aucun exploitant n'a refusé de collaborer. Cela s'est produit surtout du fait que l'ancienne structure des grosses exploitations était encore presque intacte.

Quand nous ne pensons qu'à l'estimation de la superficie, nous pouvons considérer l'enquête par segments comme plus objective et plus complète, puisqu'il n'y a pas de données manquantes et que les observations ne sont pas fondées sur les réponses des exploitants. Si nous acceptons ce principe, nous pouvons avoir une idée d'un biais possible dans le sondage agricole en faisant une comparaison avec les estimations de la superficie obtenues dans l'enquête par segments. Nous pouvons comparer, au tableau 2, les estimations pour les cultures principales de la région. Il y a une bonne correspondance entre les chiffres pour les céréales, à l'exception du blé dur, et pour les cultures permanentes, mais nous constatons pour la betterave à sucre et le soja certains problèmes qui pourraient être liés à une mauvaise compréhension de la façon de déclarer les deuxièmes récoltes au cours de la même année et le même champ ou à un biais attribuable à des valeurs manquantes. Les statistiques officielles sont produites en tenant compte de divers renseignements. Le blé dur est déclaré séparément à cause de la signification spéciale que donne à cette culture l'importante subvention accordée par la CE pour chaque hectare cultivé.

Résultats de l'enquête par segments et du sondage agricole pour les cultures principales en Emilie-Romagne (1990)

Tableau 2

Enquête par segments		Sondage agricole		ISTAT								
Emilie-Romagne	Superficie × (1,000 ha)	Superficie × (1,000 ha)	Production × (1,000 tm)	Superficie								
				Esti-mation	CV							
* Estimation corrigée au moyen d'une régression portant sur une image-satellite classée, ISTAT : Statistiques officielles. Aucun renseignement n'a été fourni sur la précision.	Blé tendre	212	5,7	208	6,9	1,177	8	212	72	38	6	
	Blé dur	46	14,9	48	15,2	260	14	72	38	6		
	Orge	43	11,2	50	17,7	184	17	38	6			
	Riz	–	–	4	59,0	23	61	6				
	Betterave à sucre	111	7,1 *	96	9,6	5,474	28	119	47	75	85	
	Soja	76	6,0 *	55	11,6	321	39	47	75	85		
	Vignes	78	13,3 *	76	18,7			75	85			
	Vergers	91	13,1 *	96	19,7			85				
	la précision.											

* Estimation corrigée au moyen d'une régression portant sur une image-satellite classée.
ISTAT: Statistiques officielles. Aucun renseignement n'a été fourni sur la précision.

Dans le sondage agricole, les coefficients de variation ont un comportement logique pour les céréales, mais ils deviennent plus difficiles à comprendre dans le cas de la betterave à sucre et du soja. Le CV (coefficient de variation) élevé pour la production peut être attribuable à des rendements plus élevés dans des exploitations agricoles plus grosses et plus spécialisées. L'estimation de la production peut être corrigée à l'aide de la différence entre les estimations de superficie obtenues dans le cadre de l'enquête par segments et du sondage agricole. Une méthode qui fait appel à un estimateur par régression pourrait être une bonne solution.

constant, mais les points relatifs à des superficies autres que des SAU et qui sont supprimés correspondent à des valeurs nulles de W_{ik} , et leur suppression peut entraîner une réduction de la variance.

4.3 Exploitations agricoles avec des champs dans différentes strates

À première vue, l'estimateur (2) semble supposer qu'une exploitation k choisie par un point dans la strate Ω_h est entièrement incluse dans cette strate. Il est évident qu'une exploitation peut avoir des champs dans différentes strates, et il faut se demander si ce fait peut avoir une incidence négative sur la fiabilité des résultats.

Rappelons que la variable utilisée n est pas réellement W_{ik} , mais X_{ik} , définie pour chaque parcelle. Le total des W ne coïncide pas avec le total des X dans chaque strate, mais c'est le cas pour toute la région pourvu que

$$\sum_i T_{ik} = A_k. \quad (7)$$

Il faut remarquer que A_k est identique à ce que nous avons déjà désigné par A_{ik} , où l'indice i n'est employé que pour indiquer que l'exploitation k a été choisie dans l'échantillon par l'intermédiaire du segment i .

Cette identité est vraie pour toute la population, quelle que soit la procédure d'échantillonnage, si les exploitations se trouvent entièrement dans la région et si la géométrie du document (la photographie aérienne) utilisé pour le relevé de terrain est exacte.

On suppose faible la perturbation attribuable aux exploitations ayant des champs dans différentes régions parce que la proportion de ces derniers est faible (en général inférieure à 1 ou 2%) et parce qu'il y a une certaine compensation du biais attribuable au fait qu'il y a des champs situés à l'intérieur de la région qui appartiennent à des exploitations dont le siège se trouve à l'extérieur de la région par le biais attribuable au fait qu'il y a des champs situés à l'extérieur de la région qui appartiennent à des exploitations dont le siège est dans la région. Nous supposons que le total de W est calculé pour les exploitations dont le siège se trouve dans la région visée par l'enquête.

4.4 Non-réponse

Nous parlons ici des estimateurs fondés sur des points correspondant à une exploitation agricole et sur des points ne correspondant pas à une exploitation agricole (section 4.1). Si un exploitant refuse de collaborer ou si l'on ne peut le retracer, on remplace la ou les rangées correspondantes à cette personne dans le tableau utilisé en entrée (tableau 1) par les valeurs moyennes pour les exploitations du segment qui ont répondu, s'il y en a; autrement, on les remplace par la moyenne des exploitations qui ont répondu pour tous les segments de la strate traitée. Si, au deuxième degré (sélection d'exploitations dans le segment), nous considérons les points correspondant à des exploitations et ceux qui ne correspondent pas à des

5. RÉSULTATS: DEUX EXEMPLES

Nous traitons ci-après de certains résultats provenant de deux régions: l'Emilie-Romagne (Italie) et la République tchèque. En République tchèque, ce sont les deux méthodes présentées respectivement dans la sous-section 2.4, 3 et dans la section 4.1 qui ont été employées; aucune donnée ne manquait. En Emilie-Romagne, le plan d'enquête général ne suivait pas exactement la procédure décrite plus haut. Les données manquantes ont été traitées selon la méthode décrite à la section 4.4.

5.1 Emilie-Romagne, 1990

En Emilie-Romagne, une superficie de 19,500 km² a été divisée en 4 strates, qui excluaient les régions montagneuses. Un échantillon de 313 segments "cadastres" (avec des limites physiques) a été tiré au moyen d'une procédure à deux degrés avec unités primaires d'échantillonnage (upé) d'environ 50 ou 100 hectares, selon la strate. On a tiré au hasard cinq points par segment à partir d'une grille ayant un pas de 50 mètres.

Des 1,565 points échantillonnés, il y en avait 326 qui ne correspondaient pas à des SAU, 206 qui correspondaient à une SAU mais pour lesquels on n'a pu trouver l'adresse de l'exploitant, 38 pour lesquels l'exploitant n'a pu être repéré et 32 où l'exploitant a refusé de collaborer. On avait des données valables pour 963 points se rapportant à des SAU et provenant de 285 segments; ces points correspondaient à 617 exploitations, dont certaines figuraient plus d'une fois dans l'échantillon.

L'introduction de valeurs pour l'exploitation agricole moyenne" mène à un biais négatif pour la variance. Pour compenser, l'exploitation n est pas incluse dans la taille de l'échantillon K_i pour le calcul des variances.

exploitations et que nous attribuons la valeur 0 aux points qui tombent sur des terres non agricoles, il est évident que l'exclusion des non-répondants introduit un biais important, parce que les valeurs nulles correspondant à des superficies autres que des SAU ne manquent jamais. Ces points ne sont pas employés pour calculer les valeurs de l'exploitation agricole moyenne" servant à remplacer les valeurs manquantes. Il y a encore un risque d'introduire un biais si les exploitants non retracés ou refusant de collaborer ont un comportement particulier, par exemple si leurs exploitations sont de façon générale plus petites ou moins efficaces.

Nous aurions pu étudier une autre façon de surmonter ce problème: l'élimination à la fois des valeurs manquantes et d'un nombre proportionnel de valeurs 0 correspondant aux points qui ne se rapportent pas à des SAU. Les deux méthodes donnent la même estimation du total, mais la seconde solution est moins commode parce que, au deuxième degré, la taille de l'échantillon n est plus un nombre entier.

La version à deux degrés de l'estimateur de Horvitz-Thompson pour le total de X dans la strate Ω_h donne :

$$X_h = \frac{1}{n_h} \sum_{i=1}^{n_h} p_i = N_h \sum_{i=1}^{n_h} \frac{1}{K_i} \sum_{k=1}^{K_i} \frac{p_{ik}}{X_{ik}} =$$

$$\left(\frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{D_i}{K_i} \sum_{k=1}^{K_i} \frac{A_{ik}}{W_{ik}} \right). \quad (2)$$

Cela signifie que, même si l'unité secondaire d'échantillonnage est la parcelle, nous n'avons pas besoin de connaître sa superficie ni X_{ik} ; nous avons seulement besoin des renseignements globaux sur l'exploitation agricole. L'estimateur est une fonction linéaire des estimations des segments sélectionnés. Sa variance dans la strate Ω_h peut être estimée par la méthode suivante (Cochran 1977, section 11.6) :

$$V(X_h) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) \sum_{i=1}^{n_h} \frac{(X_i - X_h)^2}{n_h - 1} + \frac{N_h^2}{n_h} \sum_{i=1}^{n_h} \frac{K_i(K_i - 1)}{K_i} \frac{1}{\sum_{k=1}^{K_i} \left(\frac{A_{ik}}{W_{ik} D_i} - X_i \right)^2}. \quad (3)$$

Les estimations du total sont :

$$X = \sum_H X_h = \sum_H V(X) = \sum_H V(X_h). \quad (4)$$

Les superficies cultivées sont actuellement estimées à partir de l'enquête par segments avec des données de terrain plus objectives (une observation directe faite par l'enquêteur sur le terrain), bien qu'un certain biais puisse apparaître à cause de la localisation imparfaite des segments sur le terrain. Les sondages agricoles fournissent des estimations à la fois de la superficie et de la production, mais ces estimations peuvent être entachées d'un biais plus important à cause de la non-réponse et d'une tendance subjective de l'exploitant agricole qui peut dépendre du fait qu'au moment de l'enquête ce dernier est préoccupé plus par les impôts que par les subventions ou vice-versa. La comparaison des deux estimations de la superficie, celle obtenue à partir de l'enquête par segments et celle obtenue à partir du sondage agricole, peut être utile pour vérifier si un biais n'a pas été introduit dans l'estimation de la production fondée sur le sondage agricole. Il est aussi possible d'obtenir des estimations pour les bovins, mais les résultats seront présument mauvais s'il y a un nombre élevé d'exploitations qui n'ont pas du tout de SAU et qui ne seront donc pas échantillonnées : la couverture de la base areolaire ne sera pas complète dans ce cas. Par contre, il se peut qu'il n'y ait pas de corrélation entre le nombre d'animaux d'élevage et la SAU et, par conséquent, la probabilité de sélection. Les estimations seraient alors inefficaces.

4.2 Estimation fondée seulement sur des exploitations agricoles correspondant à des points

On a été écrit pour les ordinateurs personnels (PC) (Dicorato 1993) un programme en langage C qui permet de calculer des estimations à l'aide de cette méthode. La principale partie du programme a d'abord été écrite pour calculer des estimations dans le cadre d'une enquête par segments.

Nous allons voir une autre option, où l'on utilise seulement des points qui tombent sur la SAU. Dans ce cas, nous commençons par déterminer F_i , c'est-à-dire le nombre de points qui tombent sur la SAU. (Il arrive souvent que $F_i = F_h$, constante dans chaque strate). Dans le segment i , nous observons autant de points qu'il le faut pour obtenir F_i points dans la SAU. Si le segment i ne comprend pas de SAU, nous ajoutons une observation (une exploitation agricole fictive) avec valeurs 0. Cela constitue en fait une stratification implicite au second degré ou une stratification des unités primaires (segments) en deux strates : SAU et superficies qui ne sont pas des SAU. La strate des superficies qui ne sont pas des SAU n'est pas échantillonnée. Dans ce cas, les équations (2) et (3) doivent être adaptées en remplaçant K_i par F_i et D_i par U_i . Il se peut que certaines incohérences surgissent dans les régions vallonnées parce que A_{ik} provient de la déclaration de l'exploitant agricole et U_i de segments tracés sur des photographies aériennes.

$$V(X_h) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) \sum_{i=1}^{n_h} \frac{(X_i - X_h)^2}{n_h - 1} + \frac{N_h^2}{n_h} \sum_{i=1}^{n_h} \frac{F_i(F_i - 1)}{\sum_{k=1}^{F_i} \left(\frac{A_{ik}}{W_{ik} U_i} - X_i \right)^2}. \quad (5)$$

Le deuxième membre de l'équation (6) est nul pour des segments qui n'ont pas de SAU. Ce membre ne peut être calculé si $F_i = 1$ à cause de la non-réponse. On peut attribuer une valeur 0, bien que cela amènera une sous-estimation de la variance à l'intérieur des segments, qui est relativement faible selon les calculs faits à partir des données disponibles (Carfagna 1992). Cette méthode n'a été utilisée qu'une fois pour résoudre un problème lié à la mauvaise compréhension d'instructions relatives au travail sur le terrain, qui aurait dû être effectué d'après la méthode décrite à la section 4.1. Toutefois, les avantages et les inconvénients des deux méthodes ne sont pas évidents et, jusqu'ici, aucune comparaison systématique n'a été faite pour la même région et la même année. Le fait de n'utiliser que des points correspondant à des exploitations agricoles peut accroître le coût de l'enquête si le nombre de points par segment doit être maintenu



Figure 6. Segments avec un gabarit de 5 points pour la sélection d'exploitations agricoles.

On trouve les exploitants et on leur demande de fournir des données globales sur l'exploitation agricole, y compris sur la superficie totale consacrée à chaque culture visée par l'enquête et sur la production correspondante. On ne pose pas de question sur la production de chaque champ ou de l'ensemble des champs qui se trouvent dans le segment. Cela n'est pas nécessaire parce que, dans les formules finales utilisées pour calculer les estimations (formules 2 et 3 de la section 4.1), on n'a pas besoin de la superficie cultivée ou de la production agricole de la parcelle de terre. Les instructions relatives à l'enquête sur le terrain sont habituellement transmises du CCR aux administrations nationales. Ces dernières expliquent les instructions aux coordonnateurs régionaux, qui donnent les renseignements aux enquêteurs. Les instructions peuvent être modifiées à certaines de ces étapes. Il est parfois difficile de s'assurer que les instructions n'ont pas été mal comprises, en partie du fait que les différences linguistiques sont un obstacle sérieux à la communication directe avec les enquêteurs. Dans certains pays (par exemple en Espagne), les exploitants agricoles vivent principalement dans des centres urbanisés assez grands et sont difficiles à trouver; ce qui fait qu'il peut manquer beaucoup de données.

4. ESTIMATIONS FONDÉES SUR DES EXPLOITATIONS AGRICOLES ÉCHANTILLONNÉES PAR POINTS

Nous supposons que la population Ω des segments est divisée en strates Ω_h , $h = 1, \dots, H$, que la taille totale de la population est de N segments (N_h pour la strate Ω_h) et que la taille de l'échantillon est de n segments (n_h). La taille de notre échantillon de points dans chaque segment sera K_i , déjà fixée; en général, nous avons $K_i = K$, qui

est une valeur constante pour toutes les strates; de ce nombre F_i correspondent aux exploitations agricoles sur lesquelles tombent ces points. Chaque segment i a une SAU totale U_i . Nous utilisons un plan de sondage à deux degrés. Au premier degré, le segment i est choisi avec probabilité $p_i = 1/N_h$ dans chacun des n_h essais. Au deuxième degré, l'unité n'est pas l'exploitation agricole mais la parcelle (SAU dans un segment appartenant à la même exploitation). La parcelle k du segment i a une superficie T_{ik} . La SAU totale de l'exploitation est A_{ik} pour tous les segments. U_i est la somme de la superficie T_{ik} des parcelles dans le segment i .

La méthode présentée ci-après ressemble beaucoup à la méthode dite de l'"estimateur pondéré pour un segment" utilisée aux États-Unis et au Canada (Nealon 1984).

4.1 Estimations fondées sur des points correspondant à une exploitation agricole et sur des points ne correspondant pas à une exploitation agricole

Il y aura $K - F_i$ observations (exploitations agricoles fictives) avec valeur 0 correspondant aux points qui se trouvent à l'extérieur de la SAU.

L'échantillonnage par points signifie que les parcelles sont choisies avec remise et avec une probabilité p_{ik} proportionnelle à la superficie T_{ik}/D_i (il n'est pas nécessaire de connaître la valeur de T_{ik}), où D_i est la taille du segment déterminée par la conception de la base de sondage. Nous supposons implicitement que la région étudiée est plane. Il peut s'introduire un léger biais du fait que les cultures annuelles sont habituellement exploitées sur des terres qui ne sont pas parfaitement planes et que les pâturages et les régions ne correspondant pas à une SAU se trouvent souvent sur des terrains avec une pente plus abrupte.

L'échantillonnage est effectué avec remise: une exploitation peut être choisie plus d'une fois, ce qui donne des formules plus simples pour estimer la variance. À proprement parler, la probabilité de sélection conjointe des exploitations k et k' dans l'échantillon est $p_{ikk'} \neq p_{ikk} \times p_{ikk'}$ comme cela serait le cas si l'on tirait de façon indépendante les différents points du gabarit, puisqu'il y a habituellement une distance relativement grande entre ces points. Nous ne tiendrons pas compte de ce fait ici.

W_{ik} sera une quantité additive pour une exploitation agricole, le plus souvent la production ou la superficie d'une culture particulière. Il est évident que le rendement n'est pas une variable additive. Comme nous n'avons pas de renseignements sur la distribution de W_{ik} dans l'exploitation agricole, nous créons une variable fictive X qui est uniformément distribuée et qui a, par définition, le même total que W pour chaque exploitation:

$$X_{ik} = \frac{T_{ik}}{A_{ik}} W_{ik}. \quad (1)$$

L'estimation du total de X et du total de W sont des problèmes équivalents.

à la distance entre les segments pour éviter de choisir des segments trop rapprochés. Ici, on peut utiliser des estimateurs par grappe plutôt que les formules normales pour l'échantillonnage aléatoire (Fuentes 1994; Ambrosio 1993). Avec l'échantillonnage systématique, on risque d'introduire un biais s'il existe un effet cyclique dans le paysage avec une période qui coïncide avec la taille du bloc (10 km dans l'exemple), mais cela est très peu probable. Le seuil de distance entre les segments peut amener une surestimation des erreurs types si la corrélation géométrique a une valeur positive élevée pour des distances inférieures au seuil.

La taille des segments varie d'une région à l'autre selon le paysage agricole, en particulier selon la superficie des champs. Dans la République tchèque, la superficie d'un segment était de 400 ha. Pour l'enquête areolaire, des enquêteurs repèrent les segments, tracent les champs sur une feuille transparente placée sur une photographie aérienne et notent l'utilisation du terrain. Des surveillants visitent de nouveau 5 à 10% des segments pour s'assurer qu'il n'y a pas eu d'erreur dans le travail sur le terrain. On n'utilise d'images-satellites ni pour l'enquête proprement dite ni pour le sondage agricole, mais ces images peuvent être employées pour améliorer la précision des estimations des superficies, comme nous allons le voir dans la prochaine section.

2.4.2 Amélioration des estimations de superficie au moyen d'images-satellites

On a évalué l'emploi des images-satellites à haute résolution obtenues à l'aide des capteurs TM des satellites Landsat et XS des satellites SPOT, et on les utilise encore, dans une certaine mesure, pour améliorer les estimations obtenues à partir de l'enquête sur le terrain portant sur un échantillon de segments. La méthode la plus habituelle utilise un estimateur par régression sur des images classées. On a aussi fait l'essai d'un autre estimateur basé sur des grilles de correction, et les résultats sont très proches de ceux obtenus pour l'estimateur par régression (Hay 1988; Gallego 1994). Les conclusions de cette évaluation sont semblables à celles du Département de l'agriculture des États-Unis (Allen 1990): l'utilisation d'images-satellites pour l'estimation des superficies est opérationnelle, mais encore trop coûteuse pour le gain d'efficacité obtenu. On peut atteindre le seuil de rentabilité en améliorant l'automatisation du traitement des images, car, sur le marché européen, le coût du traitement des images à cette fin est très supérieur au coût des images elles-mêmes. Ce seuil a presque été atteint en Grèce avec les images du capteur TM des satellites Landsat. Giovacchini (1992) présente des conclusions différentes sur l'analyse des coûts.

3. ÉCHANTILLONNAGE D'EXPLOITATIONS AGRICOLES AU MOYEN DE POINTS

Pour les enquêtes agricoles dans la Communauté européenne, les exploitations agricoles sont traditionnellement choisies au moyen d'une liste (Eurostat 1991). La liste est

un recensement des exploitations qui dépassent une certaine taille. Dans de nombreux pays, un recensement agricole est effectué tous les dix ans et il est rarement sinon jamais mis à jour. Il peut donc y avoir une différence importante entre la base de sondage et la population réelle au moment de l'enquête. La situation est pire dans les pays d'Europe centrale faisant partie de l'ancien bloc oriental (la région comprise entre la Pologne et la Roumanie-Bulgarie), où le changement dans la structure de la propriété foncière est si rapide que le recensement peut ne plus exister pour les fermes privées et être trop ancien pour les coopératives. On peut facilement définir des bases areolaires sur des segments carrés quand on connaît les limites géographiques de la région. Dans plusieurs pays, on utilise aussi un sous-échantillon de ces segments pour sélectionner des exploitations à l'aide d'un gabarit de points superposé au segment. Cette méthode a fait l'objet d'expériences en Allemagne, au Portugal, en Italie (Cartagna 1991) et en Espagne, et on l'utilise maintenant couramment en Grèce, en Roumanie et en République tchèque. Le gabarit est le même pour tous les segments d'une strate et il est habituellement symétrique de façon à réduire le risque de biais attribuable à un lieu géographique particulier. Les données sont obtenues seulement pour les exploitations correspondant aux points qui tombent sur une superficie agricole utilisée (SAU).

La définition de la SAU employée pour le travail sur le terrain est adaptée à chaque système national. Les bâtiments agricoles et les pacages sont inclus dans certains pays, exclus dans d'autres. Ce qui importe, c'est que la définition utilisée soit compatible avec la définition de la colonne SAU employée pour le calcul (tableau 1).

Tableau 1
Observations produites par les points sélectionnés dans le segment de la figure 6

Segment	Point	SAU	Cultures permanentes		Superficie productive		Superficie productive		Orge
			permanentes	Superficie productive	Superficie productive	Superficie productive	Superficie productive	Superficie productive	
1	1	19	4	12	64	0	0	0	0
1	1	2	0	0	0	0	0	0	0
1	1	3	0	0	0	0	0	0	0
1	4	35	0	24	131	3	3	12	12
2	1	5
2	2	35	0	24	131	3	3	12	12

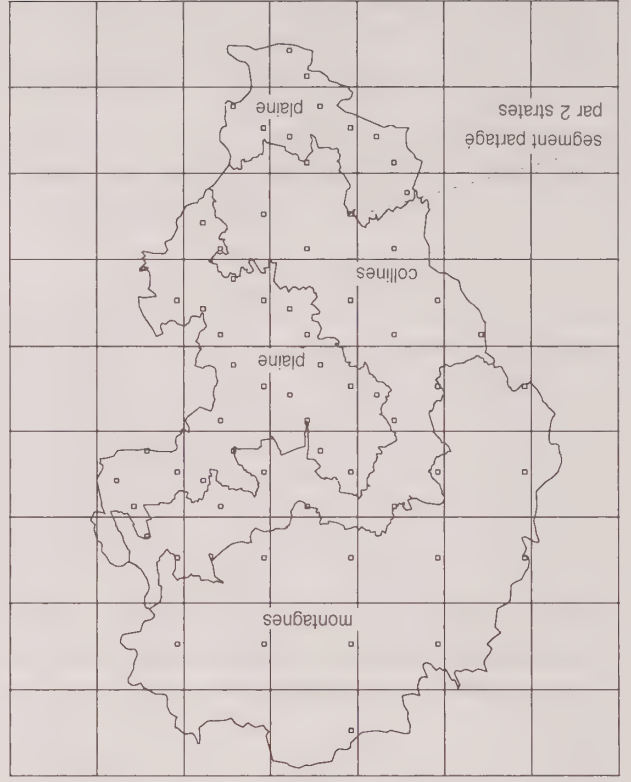
Dans l'exemple de la figure 6, le point 3 est tombé sur un terrain boisé et le point 2, sur une zone bâtie. Ces deux points produiront deux enregistrements de valeur zéro dans le fichier des exploitations agricoles. L'enquêteur devra trouver les exploitants dont les terres correspondent aux trois autres points. L'exploitation correspondant au point 1 a, dans le segment, d'autres champs qui seront inclus implicitement dans l'enquête, mais l'enquêteur n'aura pas à trouver si ces champs existent. Les points 4 et 5 apparteniront à la même exploitation, qui figurera donc deux fois dans le fichier des exploitations agricoles (tableau 1).

Depuis 1990, l'IAT a progressivement transféré l'initiative aux administrations régionales ou nationales qui souhaitent utiliser des sondages effectués à l'aide de bases aréolaires fondées sur des segments. De façon générale, les activités ont été transférées aux pays du sud de la CE et aux anciens pays communistes d'Europe centrale, qui ont manifesté beaucoup d'intérêt pour cette méthode (figure 3). Dans certains cas, comme en Italie, il n'y a qu'un échange de points de vue entre le projet national et l'IAT.

2.4.1 Prélèvement d'un échantillon de segments sur une grille carrée

Il existe deux façons principales de constituer une base aréolaire fondée sur des segments: on peut dessiner les segments sur des cartes topographiques ou cadastrales en suivant les routes, les rivières ou les limites des champs (parfois appelés segments cadastraux). L'échantillon est habituellement tiré au moyen d'une procédure à deux degrés avec unités primaires d'échantillonnage intermédiaires pour réduire le fardeau que représente l'établissement de la base de sondage (Cotter 1987), travail qui demeure, dans tous les cas, une opération lourde. Nous utilisons habituellement des bases aréolaires fondées sur une grille carrée (Gallego et Delincé 1994), que l'on peut définir beaucoup plus rapidement. Nous employons en général (mais pas nécessairement) des images-satellites pour effectuer une stratification avant l'échantillonnage.

Figure 4. Exemple d'échantillon aréolaire avec segments carrés et blocs carrés.



La figure 4 donne un petit exemple de ce genre d'échantillon avec une stratification très simple et des segments de 25 ha (hectares). L'échantillonnage est systématique, avec répétition d'une configuration dans des blocs carrés. Ici, la dimension des blocs est de 10 km sur 10 km et la configuration a quatre répétitions dans la strate la plus agricole (la plaine), deux répétitions dans les collines et une dans les montagnes. Le principal inconvénient de cette méthode est le problème lié à la gestion des segments qui chevauchent la limite entre deux strates (figure 5). Trois solutions sont mises à l'essai pour résoudre ce problème: 1) adapter la stratification à la grille d'échantillonnage, 2) séparer les segments qui chevauchent les limites en pièces appartenant à différentes strates et 3) conserver seulement la plus grosse de ces pièces.

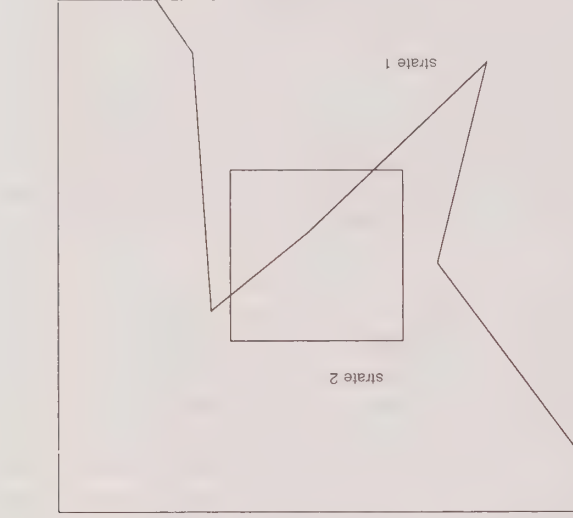
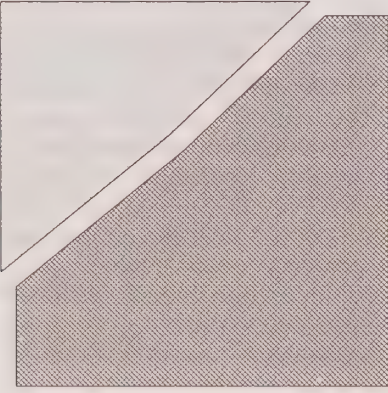


Figure 5. Un segment peut être séparé par une limite de strate.

Il n'existe pas de forte corrélation entre les erreurs non dues à l'échantillonnage les plus fréquentes – c'est-à-dire les déplacements de l'emplacement et l'inexactitude dans la forme ou la dimension du segment – et l'utilisation du terrain. On n'a pas trouvé d'influence importante sur les estimations de la superficie ou sur leur précision. La configuration d'échantillonnage à répéter dans chaque bloc est tirée au hasard, avec une restriction quant

d'environ 650 observatoires météorologiques d'Europe et des régions voisines. Ce modèle, le CGMS (Crop Growth Monitoring System ou système de surveillance de la croissance des cultures), développé en collaboration avec le WOFOST (World Food Studies Centre – ou centre mondial d'études sur les aliments – situé à Wageningen aux Pays-Bas), fait aussi appel à d'autres données, par exemple sur le sol et l'altitude, ainsi qu'à des informations sur la physiologie des plantes (van Diepen 1989; van Lanen 1992). La télédétection (images à faible résolution) sera utilisée plus tard pour l'interpolation géographique de données météorologiques observées au sol. Actuellement, on calcule les paramètres du modèle pour chaque case d'une grille de 50 km sur 50 km.

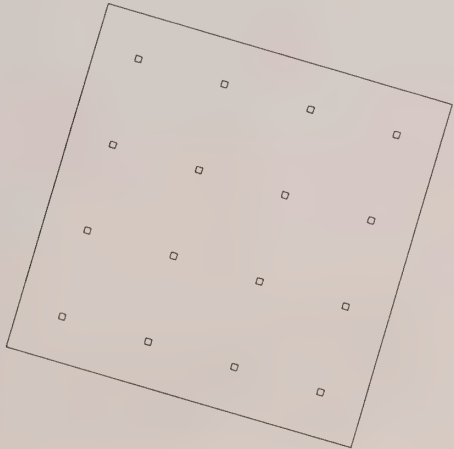
2.3 Estimations rapides au niveau de la CE

Le but principal est d'obtenir des estimations rapides de la variation de la superficie et du rendement des cultures annuelles, comparativement aux données correspondantes de l'année précédente, au moyen d'un plan de sondage à deux degrés: on utilise 53 sites (figure 1) de 40 km sur 40 km avec un échantillon de 16 segments carrés de 700 m sur 700 m (figure 2) dans chacun des sites. Les données au moyen des capteurs TM des satellites Landsat et XS des satellites SPOT. On analyse en moyenne trois images pour chaque site avec un minimum d'information de terrain, c'est-à-dire une connaissance générale des cultures dominantes dans chaque région. Un relevé de terrain est effectué en vue d'une validation a posteriori de la photointerprétation. Un rapport mensuel est produit (de mars à novembre) avec une mise à jour des estimations. Chaque rapport doit utiliser toutes les images acquises plus de 15 jours avant la production du rapport.

Figure 1. Échantillon de 53 sites pour l'estimation rapide des cultures dans la CE.



Figure 2. Segments dans un site (estimations rapides dans la CE).



2.4 Inventaires régionaux des récoltes au moyen d'un sondage par segments et de la télédétection

L'objectif de cette action était de mettre en oeuvre, d'adapter et d'évaluer des méthodes d'estimation de la superficie cultivée et de la production agricole fondées sur un échantillonnage effectué à partir d'une base aréolaire et sur des images-satellites. Quand l'IAT a commencé en 1988 à mettre cette action en oeuvre sur cinq zones pilotes d'environ 20,000 km² chacune, la priorité a été accordée aux cultures annuelles: blé tendre, blé dur, orge, colza, légumineuses à grain, tournesol, maïs, coton, tabac, betterave à sucre, pommes de terre, riz et soja, ainsi qu'aux jachères. On accorde de plus en plus d'attention aux cultures permanentes, aux pâturages et aux utilisations non agricoles des terres.

Figure 3. Régions européennes dans lesquelles des enquêtes par segments ont été réalisées en 1992.



Echantillonnage à deux degrés dans des bases aréolaires sur des segments carrés pour des sondages agricoles

F.J. GALLEGO, J. DELINCE et E. CARFAGNA¹

RÉSUMÉ

Dans le projet MARS (Monitoring Agriculture with Remote Sensing ou surveillance de l'agriculture par télédétection) de la CE (Communauté européenne), on utilise des bases aréolaires définies à l'aide d'une grille carrée pour estimer les superficies au moyen d'enquêtes sur le terrain et d'images-satellites à haute résolution. Quoique coûteuses, ces images-satellites sont utiles pour l'estimation des superficies; leur utilisation pour l'estimation du rendement n'est pas encore opérationnelle. Pour combler cette lacune, on emploie aussi les éléments de l'échantillon (segments) de l'enquête aréolaire pour échantillonner les exploitations agricoles à l'aide d'un gabarit de points superposé au segment. Le plus souvent, on utilise un nombre fixe de points par segment. On demande aux exploitants agricoles de fournir des données globales pour leur exploitation et on calcule les estimations à l'aide d'une approche de type Horvitz-Thompson. Les principaux problèmes sont la difficulté de repérer les exploitants et la vérification des cas où les instructions ont été mal comprises. On obtient de bons résultats pour la superficie et pour la production des cultures principales. Les bases aréolaires doivent être complétées par des listes (bases de sondage multiples) pour donner des estimations fiables dans le cas des animaux d'élevage.

MOTS CLÉS: Base aréolaire; sondage par points; sondage par segments; sondage d'exploitations agricoles.

1. INTRODUCTION

L'objet principal de cet article est de présenter la méthode utilisée dans le cadre du projet MARS (Monitoring Agriculture with Remote Sensing ou surveillance de l'agriculture par télédétection) de la Communauté européenne (CE) pour prélever un échantillon d'exploitations agricoles dans une base aréolaire. Dans ce projet, le prélèvement d'un échantillon d'exploitations n'est pas une activité essentielle, mais plutôt une façon d'éviter le problème de la capacité limitée des images-satellites, en particulier pour l'estimation du rendement. Nous allons d'abord présenter brièvement le projet MARS lui-même, étant donné que très peu d'articles en parlent dans les périodiques statistiques (Ambrosio 1993; Gallego 1992). On peut en trouver d'autres présentations dans des communications présentées lors de conférences (Meyer Roux 1990; Delince 1990; Shartman et coll. 1992; Carfagna et coll. 1994) ou dans des périodiques traitant de télédétection (Gonzalez et coll. 1991; Gallego et coll. 1993).

2. LE PROJET MARS DE LA COMMUNAUTÉ EUROPÉENNE

Le projet MARS a été lancé en 1988 pour évaluer et développer des applications opérationnelles de la télédétection dans le domaine des statistiques agricoles. Ce projet est réalisé par l'Institut des applications de la télédétection (IAT) du Centre commun de recherche (CCR) de la CE. La plupart des activités réalisées de 1988 à 1993 ont été divisées en quatre parties principales nommées "actions":

On élabore actuellement des modèles généraux et propres aux cultures à partir des données provenant d'un réseau

2.2 Modèles agrométéorologiques

Dans cette action, on s'intéresse aux images-satellites à faible résolution obtenues à l'aide du AVHRR (Advanced Very High Resolution Radiometer ou radiomètre perfectionné à très haute résolution) des satellites de la NOAA. Sur ces images, chaque pixel couvre une superficie d'environ 1 km² en visée verticale. Les principaux objectifs sont l'élaboration d'un logiciel convivial pour le prêt-à-l'emploi de ces images et la constitution d'une banque de données de séries chronologiques d'indices de végétation et d'autres indicateurs pour environ 3,000 unités de surveillance dans la CE. Ces unités de surveillance n'ont pas encore été définies de façon définitive. Elles devraient être des régions géographiques couvrant en gros de 500 à 1,000 km² avec un indice de végétation ou de coloration verte plus ou moins homogène (Houston 1984; Goward 1991).

2.1 Surveillance de la végétation

Des travaux sont aussi effectués dans d'autres domaines connexes, tels que l'échantillonnage à l'aide de bases aréolaires. Nous nous intéresserons ici à une méthode d'échantillonnage utilisée dans le cadre de l'action 1 "inventaires régionaux", mais nous allons d'abord parler brièvement des autres actions.

- 1) Inventaires régionaux des cultures.
- 2) Surveillance de la végétation.
- 3) Modèles agrométéorologiques.
- 4) Estimations rapides au niveau de la CE.

¹ F.J. Gallego, Centre commun de recherche des Communautés européennes, tp. 440, 21020 Ispra (Varèse) Italie; J. Delince, Commission européenne DG VI, Loi 120, 4-23/1049 Bruxelles, Belgique; E. Carfagna, Département de Statistiques, Université de Bologne, V. Belle Arti 41, 40126 Bologne, Italie.

- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAUTORY, O. (1991). La macro SAS: CALMAR. Document non publié, Institut national de la statistique et des études économiques, Paris.
- STATISTIQUE CANADA (1980). *Classification type des industries*. N° 12-501F au catalogue, Statistique Canada.
- WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.
- WU, C.F.J., et DENG, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. Dans Box, G.E.P. et coll. (Eds.), *Scientific Inference, Data Analysis and Robustness*, New York: Academic Press, 245-277.
- SUNTER, A.B. (1986). Implicit longitudinal sampling from administrative files: A useful technique. *Journal of Official Statistics*, 2, 161-168.

REMERCIEMENTS

Les auteurs tiennent à remercier René Boyer, qui leur a fourni une version de la macro SAS CALMAR adaptée à leur étude empirique, ainsi que K. P. Srinath et Michael Hidiroglou, qui ont participé à des échanges fructueux avec eux. Ils remercient également Michael Bankier et Jean Leduc pour leurs utiles commentaires sur une version antérieure de cet article.

ANNEXE A:

DÉFINITION DE LA VARIANCE DE $Y_{\text{GREG}}(d)$ ET DE L'ESTIMATEUR DE LA VARIANCE

On peut déterminer la variance de $Y_{\text{GREG}}(d)$ au moyen de l'identité

$$V(Y_{\text{GREG}}(d)) = E_1 V_2(Y_{\text{GREG}}(d)) + V_1 E_2(Y_{\text{GREG}}(d)).$$

Premièrement, considérons la variance de l'estimateur par rapport à la seconde phase d'échantillonnage, étant donné les résultats du calage de première phase. Si le vecteur de variables auxiliaires pour la pondération de seconde phase, z , contient une variable qui prend la valeur un pour tous les déclarants (ou si l'on peut construire une combinaison linéaire de variables auxiliaires qui soit égale à un pour tous les déclarants), l'estimateur par régression généralisé peut être défini par l'expression

$$Y_{\text{GREG}}(d) = \sum_{i \in s_2} w_{1i} w_{2i} y_i(d)$$

$$= \sum_{i \in s_2 \cap v} w_{1i} (y_i(d) - z_i' b_v) / p_{2i} + \sum v' b_v.$$

Si l'on fait abstraction de la variabilité due à l'estimation des coefficients de régression dans la pondération de seconde phase, on a

$$E_1 V_2(Y_{\text{GREG}}) \approx E_1 V_2 \left(\sum_{i \in s_2} w_{1i} Q_{2i} / p_{2i} \right)$$

$$= E_1 \left(\sum_{i \in s_1} \frac{1}{1 - p_{2i}} w_{1i}^2 Q_{2i}^2 \right).$$

D'après l'estimateur de la variance des estimateurs de calage proposés par Deville et Särndal (1992, p. 380), on peut formuler l'estimateur de $E_1 V_2(Y_{\text{GREG}}(d))$ de la façon suivante:

$$s_1 = \sum_{i \in s_2} \frac{1}{(1 - p_{2i})^2} (g_{1i} g_{2i} q_{2i})^2.$$

Si l'on fait abstraction de la variabilité due à l'estimation des coefficients de régression dans la pondération de première phase, le deuxième terme de l'expression de la variance peut s'écrire

BIBLIOGRAPHIE

- ARMSTRONG, J., BLOCK, C., et SRINATH, K. P. (1993). Two-phase sampling of tax records for business surveys. *Journal of Business and Economic Statistics*, 11, 407-416.
- BANKIER, M., RATHWELL, S., et MAJAKOWSKI, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census of Population. *Statistics Sweden, Workshop on the Uses of Auxiliary Information in Surveys*.
- BREWER, K. R. W., EARLY, L. J., et JOYCE, S. F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.
- CHOUDHRY, G. H., LAVALLÉE, P., et HIDIROGLOU, M. (1989). Two-phase sample design for tax data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 646-651.
- DEMING, W. E., et STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 34, 911-934.
- DEVILLE, J. C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- HIDIROGLOU, M. A., SÄRNDAL, C.-E., et BINDER, D. A. (1993). Pondération et estimation dans les enquêtes-entreprises. Présenté en décembre 1993, à l'Institut national de la statistique et des études économiques, Paris. Journées de méthodologie statistique.
- LEMAITRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- RAO, J. N. K. (1968a). Some nonresponse sampling theory when the frame contains an unknown amount of duplication. *Journal of the American Statistical Association*, 63, 87-90.
- RAO, J. N. K. (1968b). Some small sample results in ratio and regression estimation. *Journal of the Indian Statistical Association*, 6, 160-168.
- ROYALL, R. M., et EBERHARDT, K. R. (1975). Variance estimates for the ratio estimator. *Sankhyā, Série C*, 37, 43-52.
- SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Tableau 3
Comparaison de l'estimateur GREG-R2 et de l'estimateur GREG-HDP pour les dépenses totales, coefficients de variation estimés

Type de domaine	Gains engendrés par l'utilisation de GREG-R2	Aucune différence	Nombre	Nombre	
				Moyenne	Moyenne
CTI2	38	0.993	26	158	1.002
CTI3	58	0.991	40	60	1.009
CTI4	88	0.988	439		

Tableau 4
Comparaison de l'estimateur GREG-R1R2 et de l'estimateur GREG-HDP pour les dépenses totales, coefficients de variation estimés

Type de domaine	Gains engendrés par l'utilisation de GREG-R1R2	Pertes engendrées par l'utilisation de GREG-R1R2	Nombre	Nombre	
				Moyenne	Moyenne
CTI2	51	0.867	26	1.170	
CTI3	160	0.934	96	1.093	
CTI4	377	0.954	210	1.074	

Les résultats des tableaux 3 et 4 indiquent que nous serions mal avisés de remplacer l'estimateur GREG-HDP en usage actuellement par l'estimateur GREG-R1R2, malgré les perspectives intéressantes que celui-ci laisse entrevoir. Les améliorations que l'on observe avec GREG-R1R2 sont relativement négligeables, compte tenu de la forte corrélation entre le revenu du déclarant et les dépenses totales. Ces améliorations pourraient être plus notables 1) s'il y avait toujours une concordance entre les codes CTI utilisés pour la stratification de première et de seconde phase et les codes CTI qui servent à déterminer l'effectif de domaine pour des unités échantillonnées et 2) s'il n'était pas nécessaire de regrouper des strates d'échantillonnage pour faire en sorte que chaque strate a posteriori de première et de seconde phase contienne un nombre minimum d'unités.

Tableau 5
Comparaison de l'estimateur GREG-R1R2 et de l'estimateur GREG-HDP pour les dépenses totales, coefficients de variation estimés, aucune erreur de classement

Type de domaine	Gains engendrés par l'utilisation de GREG-R1R2	Pertes engendrées par l'utilisation de GREG-R1R2	Nombre	Nombre	
				Moyenne	Moyenne
CTI2	66	0.778	11	1.057	
CTI3	184	0.916	72	1.047	
CTI4	402	0.944	185	1.034	

Les résultats du tableau 5 ont été obtenus après que l'on ait modifié, pour certaines unités échantillonnées, le code CTI attribué par Revenu Canada ainsi que le code CTI utilisé pour la stratification de l'échantillon de seconde phase, cela afin d'éliminer toute incohérence entre ces codes et ceux qui servent à déterminer l'effectif des domaines. Une comparaison des tableaux 4 et 5 montre que l'efficacité relative de GREG-R1R2 est beaucoup plus grande lorsqu'il n'y a pas d'erreur de classement. Avec l'estimateur GREG-R1R2, le CV estimé diminue de plus de 22% (en moyenne) pour plus de 85% des domaines CTI2. Les résultats empiriques présentés dans cette section montrent invariablement que les gains d'efficacité réalisés grâce à l'utilisation d'information supplémentaire s'accroissent à mesure qu'augmente la taille du domaine. Cette constatation rejoint les observations faites dans la section 2 au sujet des conditions dans lesquelles la corrélation entre $y(d)$ et les vecteurs de variables auxiliaires, x et z , sera élevée. Pourvu que la variable étudiée soit fortement corrélée avec les variables auxiliaires, les corrélations qui impliquent $y(d)$ seront le plus élevées si chaque strate a posteriori qui contient au moins une unité échantillonnée du domaine d ne contient pas trop d'unités ne faisant pas partie du domaine d .

5. CONCLUSIONS

L'estimation par régression généralisée offre un cadre propice à l'utilisation d'information supplémentaire. Dans cet article, nous avons déterminé un estimateur par régression généralisée pour un plan d'échantillonnage à deux phases avec échantillonnage de Poisson à chaque phase. Nous avons étudié l'efficacité de cet estimateur en l'appliquant à l'échantillon à deux phases de dossiers fiscaux prélevé par Statistique Canada en vue d'établir des estimations annuelles de la production économique des petites entreprises. La méthode d'estimation actuellement en usage dans ce programme consiste notamment à appliquer des coefficients de correction aux estimateurs de strate formée a posteriori durant la pondération des échantillons de première et de seconde phase pour tenir compte de la différence entre la taille réelle et la taille prévue de l'échantillon. Cet estimateur de stratification a posteriori est un cas particulier de l'estimateur par régression généralisée.

Selon notre étude empirique, l'estimateur par régression (GREG-HDP) est beaucoup plus efficace que l'estimateur de Horvitz-Thompson. Notre étude compare aussi GREG-HDP à deux autres estimateurs par régression généralisés. L'utilisation de ces deux estimateurs entraîne des gains d'efficacité pour les grands domaines. Toutefois, en ce qui a trait aux petits domaines, qui intéressent particulièrement les utilisateurs d'estimations basées sur l'échantillon à deux phases de dossiers fiscaux, le rendement de ces estimateurs ne justifie pas qu'on les substitue entièrement à l'estimateur actuellement en usage.

Tableau 1
 Comparaison de l'estimateur GREG-HDP et de l'estimateur H-T pour le revenu transcrit, coefficients de variation estimés

Type de domaine	Gains engendrés par l'utilisation de GREG-HDP	Nombre		Pertes engendrées par l'utilisation de GREG-HDP	Moyenne
		Nombre	Moyenne	Nombre	Moyenne
CT12	57	0.768	20	1.113	
CT13	175	0.909	81	1.082	
CT14	359	0.945	228	1.079	

Tableau 2
 Comparaison de l'estimateur GREG-HDP et de l'estimateur H-T pour les dépenses totales, coefficients de variation estimés

Type de domaine	Gains engendrés par l'utilisation de GREG-HDP	Nombre		Pertes engendrées par l'utilisation de GREG-HDP	Moyenne
		Nombre	Moyenne	Nombre	Moyenne
CT12	57	0.773	20	1.100	
CT13	175	0.910	81	1.082	
CT14	355	0.945	232	1.079	

Canada ou qu'il ne soient pas utilisables parce qu'ils ne contiennent pas les états financiers voulus. À supposer que l'on n'ait pas à tenir compte de ces cas de non-réponse, l'estimateur GREG-HDP s'impose comme un moyen de compensation de la non-réponse.

Les résultats des tableaux 1 et 2 montrent que l'efficacité relative des estimateurs GREG-HDP et H-T est semblable pour l'une et l'autre variables étudiées. De même, les résultats des autres comparaisons faites dans le cadre de cette étude empirique varient peu selon la variable étudiée. Par conséquent, le reste des tableaux ne donne que les résultats relatifs aux dépenses totales.

Les tableaux 3 et 4 servent à comparer l'estimateur GREG-HDP aux estimateurs GREG-R2 et GREG-R1R2 respectivement. D'après les coefficients de variation estimés, GREG-R2 est un peu plus efficace que GREG-HDP. Comme une forte proportion des unités de l'échantillon de seconde phase ont une probabilité de sélection égale à un et que les deux estimateurs comparés, GREG-R2 et GREG-HDP, utilisent les mêmes variables auxiliaires dans la pondération de première phase, il n'est pas étonnant de constater le peu de différence entre les deux estimateurs. Le CV estimé de l'estimation GREG-R1R2 est généralement moins élevé que celui de l'estimation GREG-HDP et l'efficacité relative de GREG-R1R2 s'accroît à mesure qu'augmente la taille du domaine. Néanmoins, GREG-R1R2 n'est supérieure à GREG-HDP que pour 64% des domaines CT14, et l'augmentation moyenne du CV estimé pour les domaines où GREG-R1R2 est moins efficace que GREG-HDP est plus forte que la diminution moyenne du CV estimé pour les domaines où GREG-R1R2 est plus efficace.

première phase était constituée d'une ou de plusieurs strates d'échantillonnage de première phase qui avaient servi à l'échantillonnage des déclarations produites pour 1989. Ces strates étaient définies selon cinq catégories de revenu. Les strates contenues dans une strate à postériori de première phase en particulier correspondaient toutes à la même catégorie de revenu. Chaque strate à postériori de première phase devait contenir au moins vingt unités échantillonnées, cela à cause de la possibilité d'un biais dans $V(R_{GREG}(d))$ lorsque le nombre d'unités échantillonnées servant à l'estimation des coefficients de régression est très petit (Rao 1968b). Si une strate d'échantillonnage de première phase contenait moins de vingt unités échantillonnées, elle était combinée à d'autres strates correspondantes aux mêmes codes CT12 et à la même catégorie de revenu, jusqu'à ce qu'on obtienne une strate à postériori contenant au moins vingt unités échantillonnées. Nous avons constitué ainsi 166 strates à postériori de première phase. Les strates à postériori de seconde phase ont été formées de la même façon, c'est-à-dire en combinant des strates d'échantillonnage ayant trait aux mêmes codes CT14 jusqu'à ce que chaque strate à postériori compte au moins vingt unités échantillonnées. Nous avons obtenu ainsi 30 strates à postériori.

Nous avons calculé les poids de première et de seconde phase pour $R_{GREG-HDP}(d)$, $R_{GREG-R2}(d)$ et $R_{GREG-R1R2}(d)$ à l'aide d'une version modifiée de la macro SAS CALMAR (Sautory, 1991). L'ensemble de poids d'échantillonnage de première phase calculés pour l'estimateur GREG-R1R2 comprenait douze poids négatifs. En revanche, nous n'avons calculé aucun poids négatif en seconde phase pour GREG-R2 ou GREG-R1R2. (On ne peut avoir de poids négatifs avec l'estimateur GREG-HDP.) En nous servant des trois estimateurs GREG et de $R_{H-T}(d)$, nous avons calculé des estimations du revenu transcrit et des dépenses totales pour 77 domaines définis par un code CT12, 256 domaines définis par un code CT13 et 587 domaines définis par un code CT14. Comme l'estimateur GREG-R1R2 n'a produit aucune valeur négative, nous n'avons pas tenté de modifier les poids négatifs rattachés à cet estimateur. Les tableaux 1 et 2 donnent les résultats de la comparaison de l'estimateur GREG-HDP et de l'estimateur de Horvitz-Thompson (H-T). La moyenne indiquée pour les gains et les pertes est une moyenne de rapports de coefficients de variation. On constate que l'estimateur GREG-HDP est plus efficace que l'estimateur H-T pour la majorité des domaines. Sa supériorité est notable particulièrement dans le cas des domaines CT12. En ce qui concerne les domaines CT14, le coefficient de variation (CV) estimé de l'estimation GREG-HDP des dépenses totales est inférieur au CV estimé de l'estimation H-T pour 60,5% des domaines, et l'écart entre les deux CV estimés est en moyenne de 5,5%. Lorsque le CV estimé de l'estimation GREG-HDP est supérieur au CV estimé de l'estimation H-T, l'écart est de 7,9% en moyenne. Outre l'information contenue dans les tableaux 1 et 2, il y a une autre raison de préférer GREG-HDP à l'estimateur H-T: chaque année, il arrive que les dossiers fiscaux de certains déclarants de l'échantillon ne parviennent pas à Statistique

4. ÉTUDE EMPIRIQUE

Afin de comparer le rendement de $Y^{H-T}(d)$, $Y(d)$ et

$Y_{\text{GREG}}(d)$, nous avons effectué une étude empirique à l'aide de données de la province de Québec pour l'année d'imposition 1989. Comme l'estimateur $Y(d)$ est un cas particulier de $Y_{\text{GREG}}(d)$, on l'appellera $Y_{\text{GREG-HDP}}(d)$ dans l'analyse qui va suivre. (HDP est le sigle utilisé pour "Häjek deux phases".) Deux autres estimateurs par régression généralisés ont été considérés. Dans les deux cas, x et z contiennent une variable qui prend la valeur un pour tous les déclarants. L'un des deux estimateurs implique un calage par rapport au revenu du déclarant dans la pondération de seconde phase. (Le revenu du déclarant est inclus dans z comme deuxième variable auxiliaire.) L'autre estimateur implique un calage par rapport au revenu du déclarant dans les deux phases de pondération. (Le revenu du déclarant est inclus dans z comme deuxième variable auxiliaire.) Dans l'analyse qui suit, nous désignerons les estimations de totaux de domaine calculées à l'aide de ces deux estimateurs par $Y_{\text{GREG-R2}}(d)$ et $Y_{\text{GREG-R1R2}}(d)$ respectivement.

Des estimations ont été calculées pour deux variables en particulier: le revenu transcrit et les dépenses totales. Il y a quelques différences conceptuelles entre le revenu transcrit et le revenu du déclarant. Par exemple, dans de nombreuses industries les gains en capital et les produits exceptionnels sont inclus dans le revenu du déclarant mais exclus du revenu transcrit. En outre, le revenu du déclarant donne lieu plus souvent à des erreurs de saisie que le revenu transcrit étant donné qu'il n'est pas soumis aux mêmes normes de contrôle qualitatif.

L'univers étudié comprenait environ 140,000 déclarants T2 qui avaient indiqué des revenus de plus de 25,000\$ pour l'année d'imposition 1989. Nous avons appliqué les probabilités d'échantillonnage de première et de seconde phase qui avaient servi à l'échantillonnage de déclarations produites pour l'année d'imposition 1989. L'échantillon de première phase comptait environ 31,000 déclarants tandis que l'échantillon de seconde phase comprenait environ 23,000 entreprises. La corrélation entre le revenu du déclarant et le revenu transcrit pour les entreprises de l'échantillon de seconde phase était de 0,969, tandis que la corrélation entre le revenu du déclarant et les dépenses totales était de 0,960. Une forte proportion des unités des échantillons de première et de seconde phase ont été échantillonnées avec une probabilité égale à un. Toutes les unités ayant une probabilité de sélection égale à un dans la première phase ont été exclues de la pondération de première phase et la valeur des poids g correspondants a été fixée à un. Les unités qui avaient une probabilité de sélection égale à un dans la seconde phase ont été traitées de la même manière dans la pondération de seconde phase. L'échantillon de première phase comptait 9,884 unités dont la probabilité de sélection en première phase était différente de un, tandis que l'échantillon de seconde phase comptait 910 unités dont la probabilité de sélection en seconde phase était elle aussi différente de un. Chaque strate a posteriori de

Choudhry, Lavallée et Hidiroglou (1989) notent que la variance de $Y(d)$ est définie approximativement par l'expression

$$V(Y(d)) \approx \sum_{i=1}^n \sum_{j=1}^n \left(\frac{1 - p_{ij}}{p_{ij}} \right) Y_i(d) Y_j(d) - \left(\frac{1 - p_{ii}}{p_{ii}} \right) Y_i(d)^2$$

où $Y_u(d)$ et $Y_v(d)$ sont les totaux de la variable y pour les parties du domaine d contenues dans les strates u et v respectivement.

L'estimateur de cette variance est défini par l'expression

$$V(Y(d)) = \sum_{i=1}^n \sum_{j=1}^n \left(\frac{1 - p_{ij}}{p_{ij}} \right) \left(\frac{N_u}{N_v} \right)^2 Y_i(d) Y_j(d) - \left(\frac{1 - p_{ii}}{p_{ii}} \right) \left(\frac{N_u}{N_v} \right)^2 Y_i(d)^2$$

où les estimations N_u et N_v sont calculées à l'aide des poids finals.

La présence du facteur $(N_u/N_v)^2$ peut s'expliquer par l'amélioration des propriétés conditionnelles de l'estimateur (Royall et Eberhardt 1975). Wu (1982) a lui aussi étudié un estimateur de variance qui contient un facteur de correction analogue; il s'agit de l'estimateur de la variance d'un estimateur par quotient pour un échantillonnage à une phase. L'étude empirique de Wu et Deng (1983) montre que l'utilisation du facteur de correction contribue à améliorer les propriétés de couverture des intervalles de confiance basés sur l'approximation normale.

L'estimateur $Y(d)$ est un cas particulier de $Y_{\text{GREG}}(d)$, qui peut s'appliquer lorsque la pondération des échantillons de première et de seconde phase fait intervenir une seule variable auxiliaire qui a la valeur un pour tous les déclarants. Dans ce cas, nous avons $g_{1i} = N_u/N_v$ pour tous les déclarants de la strate u de la première phase et $g_{2i} = N_v/N_v$ pour tous les déclarants de la strate v de la seconde phase. Notons que le choix de ces variables auxiliaires élimine la possibilité de poids g négatifs. L'estimateur de variance $V(Y(d))$ diffère peu de l'estimateur $V(Y_{\text{GREG}}(d))$ pour ce cas particulier de $Y_{\text{GREG}}(d)$. Le poids g de seconde phase apparaît dans le premier terme de $V(Y(d))$ mais non dans $V(Y_{\text{GREG}}(d))$.

3.2 Estimateur de Horvitz-Thompson

L'échantillon de seconde phase est un échantillon

d'entreprises prélevé à l'aide d'entités statistiques. Comme certaines entreprises sont des sociétés en nom collectif, il se peut que plusieurs entités statistiques se rapportent à la même entreprise. Si l'on veut établir des estimations pour l'univers des entreprises, il faut prévoir une correction qui tienne compte de l'effet de l'existence de sociétés en nom collectif. Si l'entreprise j est une société en nom collectif, elle sera incluse dans l'échantillon de seconde phase si au moins un des déclarants qui ont un intérêt dans cette entreprise est échantillonné. Pour éviter que l'estimateur de Horvitz-Thompson habituel ne produise une surestimation, il faut introduire un facteur de correction qui tienne compte de l'existence de sociétés en nom collectif. Désignons par δ_{ij} la participation du déclarant i dans l'entreprise j et supposons que l'entité statistique (i, j) est échantillonnée dans la seconde phase. On corrige les chiffres relatifs à l'entreprise j au moyen du facteur δ_{ij} de manière que seule la part des revenus et des dépenses rattachée au déclarant i entre en ligne de compte dans les estimations. Rao (1968a) décrit une méthode de correction semblable dans un contexte un peu différent.

Soit y_j la valeur de la variable y pour l'entreprise j . L'estimateur de Horvitz-Thompson du total de y pour le domaine d , Y compris la correction pour les sociétés en nom collectif, est défini par l'expression

$$Y^{H-T}(d) = \sum_{i \in s_2} \sum_{j \in J_i} \delta_{ij} y_j (d) / (p_{1i} p_{2i}),$$

où J_i est l'ensemble des indices des entreprises qui apparaissent en tout ou en partie au déclarant i . Comme les probabilités d'échantillonnage dépendent uniquement de l'indice i , $Y^{H-T}(d)$ peut s'écrire

$$Y^{H-T}(d) = \sum_{i \in s_2} y_i (d) / (p_{1i} p_{2i}),$$

où

$$y_i (d) = \sum_{j \in J_i} \delta_{ij} y_j (d).$$

$Y^{H-T}(d)$ est un estimateur sans biais du total de y pour les entreprises du domaine d . Voir Rao (1968a).

On obtient l'échantillon de seconde phase en soumettant l'échantillon de Poisson prélevé en première phase à un nouvel échantillonnage de Poisson. L'échantillon de seconde phase est donc, lui aussi, un échantillon de Poisson et la variance de $Y^{H-T}(d)$ est

$$V(Y^{H-T}(d)) = \sum_{i=1}^I [(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})] y_i (d)^2.$$

Un estimateur non biaisé de cette variance est défini par l'expression

$$V(d) = \sum_{i \in s_2} w_i y_i (d).$$

L'estimateur de stratification a posteriori du total de y pour le domaine d est défini par l'expression

$$w_i = w_{1i} w_{2i}.$$

et le poids final est

$$w_{2i} = (1/p_{2i}) (N_v/N_v)$$

Le poids de stratification a posteriori de l'entité statistique $(i, j) \in v$, pour la seconde phase est

$$N_v = \sum_{i \in s_2 \cap v} w_{1i} / p_{2i}.$$

Un autre estimateur du nombre de déclarants inclus dans la strate a posteriori v , basé uniquement sur les unités de l'échantillon de seconde phase, est

$$\tilde{N}_v = \sum_{i \in s_1 \cap v} w_{1i}.$$

On peut estimer le nombre de déclarants inclus dans la strate a posteriori v de seconde phase, d'après l'échantillon de première phase, par l'équation

$$w_{1i} = (1/p_{1i}) (N_u/\tilde{N}_u).$$

Le poids de stratification a posteriori du déclarant i , $i \in u$, pour la première phase est

$$\tilde{N}_u = \sum_{i \in s_1 \cap u} (1/p_{1i}).$$

Des coefficients de correction sont appliqués aux estimateurs de strate formée a posteriori durant la pondération des échantillons de première et de seconde phase. Choudhry, Lavallée et Hidiroglou (1989) font une analyse générale de la pondération d'un échantillon de Poisson à deux phases avec des coefficients de correction appliqués dans un contexte de stratification a posteriori. Supposons que la strate a posteriori u de première phase contient N_u déclarants. On peut estimer le nombre de déclarants inclus dans la strate a posteriori u de première phase, d'après l'échantillon de première phase, par l'équation

Brewer, Early et Joyce (1972) ont proposé de corriger l'estimateur de Horvitz-Thompson de manière à tenir compte de la différence entre la taille réelle et la taille prévue des échantillons de Poisson. On retrouve ces corrections dans la méthode actuellement utilisée pour calculer des estimations à partir de l'échantillon à deux phases de dossiers fiscaux.

3.3 Estimateur de Horvitz-Thompson avec stratification a posteriori

$$V(Y^{H-T}(d)) = \sum_{i \in s_2} [(1 - p_{1i} p_{2i}) / (p_{1i} p_{2i})] y_i (d)^2.$$

première phase sont mises à jour chaque année. Cette opération est nécessaire pour deux raisons: i) un déclarant peut changer de strate d'échantillonnage de première phase d'une année d'imposition à l'autre; ii) les fractions de sondage de première phase pour une strate donnée peuvent varier d'une année à l'autre.

Revenu Canada expédie à Statistique Canada une copie des déclarations de revenus et des états financiers des déclarants qui font partie de l'échantillon de première phase. Pour former l'échantillon de seconde phase, on crée des entités statistiques à partir de données sur les entreprises correspondantes aux déclarants de l'échantillon de première phase. Posons $J = \{j\}$ comme l'univers des entreprises visées par l'échantillonnage de dossiers fiscaux. Une entité statistique, désignée par (i, j) , est créée pour chaque combinaison déclarant-entreprise dans l'échantillon de première phase. Pour chaque déclarant TI , les données sur les entreprises que possède entièrement ou partiellement ce déclarant (y compris son degré de participation) – et qui sont nécessaires à la création des entités statistiques – se trouvent dans les déclarations de revenus et dans les états financiers qui y sont annexés. Comme il y a une correspondance biunivoque entre les entreprises et les déclarants TI , une seule entité statistique est créée pour chaque déclarant TI de l'échantillon de première phase. Pour chaque année d'imposition, Statistique Canada attribue un code $CTI4$ aux entités statistiques qui n'ont jamais fait partie de l'échantillon. Ce code est déterminé au moyen d'informations qui ne se limitent pas à l'activité économique mentionnée sur la déclaration de revenus, et il est plus précis, pour les troisième et quatrième chiffres, que le code attribué par Revenu Canada. Pour les entités statistiques qui ont déjà fait partie de l'échantillon, elles conservent le code $CTI4$ qui leur avait été attribué.

D'un point de vue conceptuel, l'échantillon de seconde phase est un échantillon d'entreprises. D'un point de vue technique, c'est un échantillon de déclarants prélevé au moyen d'entités statistiques. Celles-ci sont stratifiées selon le code $CTI4$ attribué par Statistique Canada et selon la province et la taille. Le revenu total de l'entreprise j sert de variable de taille pour l'entité statistique (i, j) . Si une entité statistique rattachée à un déclarant TI en particulier est échantillonnée en seconde phase, alors toutes les entités statistiques qui se rapportent à ce déclarant seront échantillonnées. Par conséquent, la probabilité d'échantillonnage de seconde phase pour l'entité statistique (i, j) dépend uniquement de i .

L'échantillon de seconde phase est prélevé selon la méthode de Poisson, où l'on utilise des nombres factices générés à partir du numéro d'identification des déclarants. La fonction de calcul utilisée pour l'échantillonnage de seconde phase n'a aucun rapport avec celle qui est employée pour l'échantillonnage de première phase. Les données relatives à quelque 35 variables financières sont extraites des déclarations de revenus et des états financiers des déclarants échantillonnés dans la seconde phase. Si nécessaire, on met à jour les codes $CTI4$ attribués par Statistique Canada pour faire en sorte que tous les codes à quatre chiffres utilisés dans le calcul des estimations correspondent à l'année d'imposition pertinente.

Les déclarants qui ont des intérêts dans des entreprises sont classés par Revenu Canada selon la CTI . En règle générale, l'activité économique mentionnée sur la déclaration de revenus permet de déterminer assez précisément le code à deux chiffres du déclarant. Revenu Canada élargit ce code à trois ou à quatre chiffres pour la plupart des déclarants. Cependant, les déclarants n'ont pas tous un code à quatre chiffres, et les deux derniers chiffres du code $CTI4$ attribué par Revenu Canada sont assez imprécis. L'échantillonnage à deux phases de dossiers fiscaux a pour but de permettre une estimation plus précise de la production économique au niveau d'agrégation à quatre chiffres. La section 3.1 contient une brève description du plan d'échantillonnage à deux phases. Armstrong, Block et Srinath (1993) traitent plus en détail ce type d'échantillonnage. Les sections 3.2 et 3.3 portent sur l'estimation dans un contexte d'échantillonnage à deux phases; la section 3.2 traite de l'estimateur de Horvitz-Thompson et la section 3.3 examine un estimateur de stratification a posteriori.

3.1 Plan d'échantillonnage

Les données administratives qui servent à constituer la base de sondage pour une année d'imposition particulière sont tirées des déclarations de revenus que traite Revenu Canada sur deux années civiles. L'échantillonnage de Poisson offre donc de nets avantages opérationnels parce qu'on peut commencer l'échantillonnage avant même de disposer d'une base de sondage complète. La population visée par l'échantillonnage de dossiers fiscaux est l'univers des entreprises dont le revenu brut dépasse 25,000\$, à l'exception des grandes entreprises visées par les enquêtes postales. L'échantillon de première phase est un échantillon longitudinal de déclarants. Les strates sont définies selon le niveau à 2 chiffres de la Classification type des industries ($CTI2$), la province et la taille de l'entreprise (revenu brut). Tous les déclarants qui font partie de l'échantillon de première phase pour l'année d'imposition T et dont le dossier peut être échantillonné à nouveau pour l'année d'imposition $T + 1$ demeurent dans l'échantillon pour l'année $T + 1$. Chaque année, on peut ajouter des déclarants dans l'échantillon de première phase afin d'accroître la précision de certaines estimations et de remplacer les déclarants échantillonnés les années précédentes mais qui ne font plus partie du champ de l'enquête.

Pour effectuer l'échantillonnage de Poisson en première phase, on attribue à chaque déclarant un nombre pseudo-aléatoire (nombre factice) qui se situe dans l'intervalle $(0, 1)$. Ce nombre est généré par une fonction de calcul qui utilise le numéro d'identification du déclarant comme paramètre, puis il est comparé à l'intervalle d'échantillonnage pour la strate correspondante. Si un nombre factice particulier se situe dans l'intervalle de sondage en question et que le déclarant correspondant n'est pas déjà inclus dans l'échantillon de première phase, alors ce déclarant est ajouté à l'échantillon. Comme les identificateurs de déclarants ne changent pas, l'échantillonnage de Poisson facilite la formation d'un échantillon longitudinal de première phase. Les probabilités d'échantillonnage en première phase des déclarants qui font déjà partie de l'échantillon de

Suivant le type d'information supplémentaire qui sera utilisée, les poids g rattachés à l'estimateur par régression généralisé et, par conséquent, aux estimations correspondantes pourraient être négatifs.

3. APPLICATION: ÉCHANTILLONNAGE À DEUX PHASES DE DOSSIERS FISCAUX

L'échantillonnage à deux phases de dossiers fiscaux s'inscrit dans une stratégie globale de Statistique Canada concernant la production d'estimations annuelles sur l'activité économique au Canada. Pour les grandes entreprises, les données sont recueillies au moyen d'enquêtes postales; pour les petites, elles proviennent d'un échantillon de dossiers fiscaux. La combinaison des estimations d'enquête et des estimations fondées sur l'échantillon de dossiers fiscaux permet d'estimer la valeur de variables financières pour l'univers des entreprises. En ce qui regarde le calcul d'estimations pour les petites entreprises seulement, on préfère utiliser les données fiscales plutôt que les données d'enquête afin de réduire les coûts et le fardeau de réponse.

L'échantillonnage à deux phases a été rendu nécessaire parce qu'il fallait faire des estimations pour des domaines définis selon le code à quatre chiffres de la Classification type des industries (CTI) (Statistique Canada 1980). Les deux premiers chiffres du code (CTI2) permettent de diviser l'activité économique en 76 groupes dans un premier temps. Ensuite, les deux derniers chiffres permettent d'obtenir une classification plus détaillée à l'intérieur de chaque groupe (CTI4). Par exemple, une entreprise pourrait être classée dans l'industrie du transport d'après son code CTI2 et dans l'industrie du camionnage de vrac liquide d'après son code CTI4.

Il y a deux types de déclarants: ceux qui produisent une formule T1 (déclarants T1) et ceux qui produisent une formule T2 (déclarants T2). Les premiers sont des particuliers, qui peuvent posséder entièrement ou partiellement une ou plusieurs entreprises non constituées en société, tandis que les seconds sont des entreprises constituées en société. C'est Revenu Canada, le ministère du gouvernement canadien chargé de la perception de l'impôt, qui fournit à Statistique Canada les fichiers administratifs contenant de l'information restreinte sur tous les déclarants qui ont des intérêts dans des entreprises. Ces fichiers servent à établir une base de sondage. Cette base ne contient cependant aucune information sur le nombre d'entreprises que possèdent les déclarants T1 ni sur le degré de participation de ces personnes. Mais elle contient des données géographiques et des données sur le revenu brut et le bénéfice net d'entreprise pour les deux types de déclarants. Elle contient aussi des données sur quelques autres variables financières importantes, dont la rémunération et les stocks, pour les déclarants T2. On cherche à établir des estimations sur quelques 35 variables financières qui figurent dans les déclarations de revenus et les états financiers qui y sont annexés, mais non dans les fichiers administratifs fournis par Revenu Canada.

On peut définir l'estimateur de la variance approximative de $Y_{\text{GREG}}(d)$ par l'équation

$$V(Y_{\text{GREG}}(d)) = \sum_{i=1}^I \frac{1 - p_{1i}}{p_{1i}^2} p_{2i}^2 (g_{1i} q_{1i})^2 + \sum_{i=1}^I \frac{1 - p_{2i}}{p_{2i}^2} (g_{1i} g_{2i} q_{2i})^2.$$

Comme $y(d)$ n'est connue que pour les unités de s_2 , les estimations de B_u et B_y sont calculées au moyen des équations

$$B_u = \left(\sum_{i \in s_2 \cap u} w_i x_i x_i' \right)^{-1} \left(\sum_{i \in s_2 \cap u} w_i x_i y_i(d) \right),$$

$$B_y = \left(\sum_{i \in s_2 \cap v} w_i z_i z_i' \right)^{-1} \left(\sum_{i \in s_2 \cap v} w_i z_i y_i(d) \right).$$

Les résidus d'échantillon requis pour calculer l'estimateur de la variance sont $q_{1i} = y_i(d) - x_i' B_u$ et $q_{2i} = y_i(d) - z_i' B_y$. L'annexe A donne plus de détails sur le calcul de la variance approximative de $Y_{\text{GREG}}(d)$ et de l'estimateur de cette variance.

Si y est fortement corrélée avec x et z , la variance de l'estimateur par régression généralisé du total de population de y sera plutôt faible. Cependant, il convient de souligner que l'existence d'une forte corrélation entre y , d'une part, et x et z , d'autre part, ne signifie pas nécessairement que la variance de l'estimateur du total de y pour un domaine en particulier sera assez peu élevée, puisque $y(d)$ peut être faiblement corrélée avec x et z dans les strates a postérieur qui contiennent au moins une unité de l'échantillon appartenant au domaine d .

La corrélation de $y(d)$ avec x et z dans une strate a postérieur compte des unités échantillonnées qui n'appartiennent pas au domaine d pour être faible si la strate a postérieur compte des unités échantillonnées qui n'appartiennent pas au domaine d . Ce cas peut se produire souvent si l'on ne connaît pas les totaux de domaines pour les variables auxiliaires ou si l'on ne peut obtenir de données exactes sur l'effectif des domaines en ce qui concerne les unités de l'échantillon de première phase. Dans le cas de l'échantillonnage à deux phases pour stratification, on ne connaît rien sur l'effectif des domaines avant le tirage de l'échantillon de première phase. Si chaque strate a postérieur de première phase est formée par le regroupement de plusieurs strates d'échantillonnage de première phase, par exemple, la plupart des strates a postérieur contiendront plus d'un domaine. La variable θ utilisée pour prédire l'effectif des domaines durant la stratification de l'échantillon de première phase n'est pas un prédicteur exact. Si les strates a postérieur de seconde phase sont formées par le regroupement de strates d'échantillonnage de la seconde phase, chaque domaine peut être réparti entre un certain nombre de strates a postérieur de la seconde phase.

$U = \{u\}$ et $V = \{v\}$ les ensembles de "strates a posteriori" respectivement de première et de seconde phase. Dans la pondération par régression généralisée, on corrige les poids initiaux $1/p_{1i}$ de manière à obtenir des poids $w_{1i} = g_{1i}/p_{1i}$ qui satisfont les équations de calage

$$\sum_{i \in S1 \cap v} w_{1i} x_i = X_n,$$

pour chaque "strate a posteriori" de première phase u , où x_i est un vecteur $L_1 \times 1$ de variables auxiliaires connu pour toutes les unités de la population et X_n est le vecteur de totaux de variables auxiliaires pour la strate a posteriori u . Les poids corrigés ont pour effet de minimiser la mesure de distance $\sum_{i \in S1} (g_{1i} - 1)^2/p_{1i}$. On peut obtenir les mêmes poids avec un modèle en se servant de l'équation

$$E_{\xi}(y_i) = x_i' \beta_u, i \in u \\ V_{\xi}(y_i) = \sigma^2,$$

où y_i est la valeur de la variable étudiée pour l'unité i , et $E_{\xi}(\cdot)$ et $V_{\xi}(\cdot)$ désignent respectivement l'espérance et la variance de modèle.

En ce qui concerne les estimateurs par régression généralisés qui nous intéressent, la pondération de l'échantillon de seconde phase comprend une opération de calage qui dépend des résultats de la pondération en première phase. On corrige les poids initiaux, w_{1i}/p_{2i} , de manière à obtenir les poids finals $w_i = g_{2i}w_{1i}/p_{2i}$, qui satisfont les équations de calage

$$\sum_{i \in S2 \cap v} w_i z_i = Z_v,$$

pour chaque strate a posteriori de seconde phase v , où z_i est un vecteur $L_2 \times 1$ de variables auxiliaires connu pour toutes les unités de l'échantillon de première phase et $Z_v = \sum_{i \in S1 \cap v} w_{1i} z_i$ est l'estimation du vecteur de totaux de variables auxiliaires pour la strate a posteriori v , calculée à l'aide des poids corrigés de la première phase, w_{1i} . Notons que ces équations de calage diffèrent sensiblement des exemples contenus dans Särndal et Swensson (1987, p. 284-288) et Särndal, Swensson et Wretman (1992, p. 359-366) du fait qu'elles renforcent des poids de première phase corrigés au lieu de poids initiaux. Les poids finals ont pour effet de minimiser la mesure de distance $\sum_{i \in S2} w_{1i} (g_{2i} - 1)^2/p_{2i}$. On peut obtenir les mêmes poids avec un modèle en se servant de l'équation

$$E_{\xi}(w_i y_i) = w_i z_i' \beta_v, i \in v \\ V_{\xi}(w_i y_i) = w_i \sigma^2.$$

Il y a deux grands avantages à utiliser des poids de première phase corrigés au lieu de poids initiaux dans les équations de calage de seconde phase. Premièrement, on peut définir l'estimateur par régression généralisée pour le domaine d par l'expression

$$B_v = \left(\sum_{i \in S1 \cap v} w_{1i} z_i z_i' \right)^{-1} \left(\sum_{i \in S1 \cap v} w_{1i} z_i y_i \right) (d).$$

$$Y^{\text{GREG}}(d) = \sum_{i \in S2} y_i(d) g_{1i} g_{2i} / p_{1i} p_{2i},$$

qui utilise des "poids g_i " de première et de seconde phase. Deuxièmement, supposons que des variables auxiliaires servent au calage dans les deux phases de pondération. On peut alors utiliser les poids finals pour obtenir, pour ces variables, des estimations de totaux de population qui soient égales au total réel. Désignons par $X_n'' = \sum_{i \in S1 \cap u} x_i' / p_{1i}$ le vecteur $L_1 \times 1$ des estimations Horvitz-Thompson de totaux de variables auxiliaires pour la strate a posteriori de première phase u . Le poids g de première phase est défini par l'équation

$$g_{1i} = 1 + \lambda_n' x_i,$$

où $\lambda_n' = (X_n'' - X_n') M_n^{-1}$ et $M_n^{-1} = (\sum_{i \in S1 \cap u} x_i x_i' / p_{1i})^{-1}$. Pour la strate a posteriori de seconde phase v , désignons l'estimation de Z_v fondée sur des poids initiaux de seconde phase par $Z_v' = \sum_{i \in S2 \cap v} w_{1i} z_i' / p_{2i}$. Le poids g de seconde phase est défini par l'équation

$$g_{2i} = 1 + \lambda_v' z_i,$$

où $\lambda_v' = (Z_v' - Z_v') M_v^{-1}$ et $M_v^{-1} = (\sum_{i \in S2 \cap v} w_{1i} z_i z_i' / p_{2i})^{-1}$. La variance approximative de $Y^{\text{GREG}}(d)$ est définie par l'équation

$$V(Y^{\text{GREG}}(d)) \approx \sum_{i=1}^I \frac{1}{1 - p_{1i}} p_{1i} \bar{Q}_{2i}^2 + E_1 \left[\sum_{i \in S2} \frac{1}{1 - p_{2i}} (w_{1i} \bar{Q}_{2i})^2 \right],$$

où $E_1(\cdot)$ désigne l'espérance par rapport à la première phase d'échantillonnage, $\bar{Q}_{1i} = y_i(d) - x_i' B_u$ pour chaque unité contenue dans la strate a posteriori de première phase u , et $B_u = \left(\sum_{i \in u} x_i x_i' \right)^{-1} \left(\sum_{i \in u} x_i y_i \right) (d)$. si $y(d)$ était connue pour toutes les unités de la strate a posteriori de première phase u , et B_u , le vecteur des coefficients estimés de la régression de $y(d)$ par rapport à x que l'on obtiendrait la régression de $y(d)$ par rapport à x que l'on obtiendrait si $y(d)$ était connue pour toutes les unités de la strate a posteriori de première phase u , est défini par l'équation

$$B_u = \left(\sum_{i \in u} x_i x_i' \right)^{-1} \left(\sum_{i \in u} x_i y_i \right) (d).$$

De même, $\bar{Q}_{2i} = y_i(d) - z_i' B_v$ pour chaque unité contenue dans la strate a posteriori de seconde phase v et B_v , le vecteur des coefficients estimés de la régression de $y(d)$ par rapport à z que l'on obtiendrait - moyennant le calage de la première phase - si $y(d)$ était connue pour toutes les unités de l'échantillon de première phase contenues dans la strate a posteriori de seconde phase v , est défini par l'équation

Estimation par régression généralisée pour un échantillon à deux phases de dossiers fiscaux

JOHN ARMSTRONG et HÉLÈNE ST-JEAN¹

RÉSUMÉ

Dans cet article, nous déterminons un estimateur par régression généralisé pour domaines ainsi qu'un estimateur approximatif de la variance correspondante suivant un plan d'échantillonnage à deux phases pour stratification avec échantillonnage de Poisson à chaque phase. Ces estimateurs sont une application du modèle général d'estimation par régression pour l'échantillonnage à deux phases de dossiers fiscaux de Statistique Canada, formé annuellement. Enfin, nous comparons à l'estimateur de Horvitz-Thompson trois cas particuliers de l'estimateur par régression généralisé, soit deux estimateurs par régression et un estimateur de stratification a posteriori.

MOTS CLÉS: Estimation fondée sur un modèle; estimation pour domaines; échantillonnage de Poisson.

1. INTRODUCTION

Nous étudions dans cet article le problème de l'estimation pour domaines suivant un échantillonnage à deux phases pour stratification lorsque l'échantillonnage de Poisson est utilisé dans les deux phases. Considérons une population de N unités et supposons qu'il faut estimer le total d'un caractère étudié, y , pour L domaines disjoints. On peut prédire convenablement, mais non avec exactitude, l'effectif des domaines à l'aide d'une variable auxiliaire, θ , dont les valeurs ne sont pas observées avant l'échantillonnage. Il est moins coûteux d'obtenir de l'information sur θ que sur y et moins coûteux aussi que d'obtenir des données exactes sur l'effectif des domaines. Dans la première phase d'échantillonnage, on prélève un échantillon de Poisson dans la population et l'on observe la valeur de θ pour chaque unité échantillonnée. Les unités sont ensuite stratifiées au moyen des valeurs de θ . Cette opération ressemble à une stratification par domaine. Dans la seconde phase d'échantillonnage, on tire un échantillon de Poisson dans chaque strate et l'on observe la valeur de y pour chaque unité de l'échantillon, tout en recueillant des données exactes sur l'effectif des domaines.

L'estimateur de Horvitz-Thompson du total de y pour le domaine d est $X^{H-T}(d) = \sum_{i \in s_2} y_i(d) / (p_{1i} p_{2i})$, où $y_i(d)$ prend la valeur de y_i si l'unité i appartient au domaine d et la valeur zéro dans le cas contraire, s_2 désigne l'échantillon de seconde phase, et p_{1i} et p_{2i} désignent les probabilités de sélection de l'unité i pour les première et seconde phases, respectivement. Comme la taille des échantillons prélevés au moyen d'un échantillonnage de Poisson est une variable aléatoire, l'estimateur ci-dessus peut n'être pas efficace. (Voir Sunter 1986, ou Särndal, Swensson et Wretman 1992, p. 63.) L'estimateur par régression généralisé peut remplacer l'estimateur de

Horvitz-Thompson lorsqu'on dispose d'information supplémentaire. Dans cet article, nous déterminons un estimateur par régression généralisé pour l'échantillonnage de Poisson à deux phases ainsi qu'un estimateur approximatif de la variance correspondante.

Dans la section 2, nous établissons l'estimateur par régression généralisé et l'estimateur approximatif de la variance. Ensuite, nous dérivons dans la section 3 l'application qui a fait ressortir le problème de l'estimation, soit l'échantillon à deux phases de dossiers fiscaux de Statistique Canada formé annuellement. Enfin, dans la section 4, nous présentons les résultats d'une étude empirique où l'estimateur de Horvitz-Thompson est comparé à trois cas particuliers de l'estimateur par régression généralisé, soit l'estimateur de stratification a posteriori utilisé actuellement dans le plan de sondage appliqué aux déclarations et deux estimateurs par régression.

2. ESTIMATION PAR RÉGRESSION GÉNÉRALISÉE

L'estimation par régression généralisée n'est pas nouvelle. Deming et Stephan (1940) décrivent un estimateur par régression généralisé pour un plan d'échantillonnage à une phase. Lemaître et Dufour (1987) et Bankier, Rathwell et Majkowski (1992) décrivent des applications récentes de l'estimation par régression généralisée à Statistique Canada. Hidiroglou, Särndal et Binder (1993) traitent en profondeur l'utilisation des estimateurs par régression généralisés dans les enquêtes-entreprises.

On peut aborder la question des estimateurs par régression généralisés du point de vue de l'échantillonnage fondé sur un modèle (Särndal, Swensson et Wretman 1992) ou du point de vue du calage (Dewille et Särndal 1992). Soient

¹ John Armstrong, Division des études sociales et économiques, 24 - Immeuble R.H. Coats, Statistique Canada, Parc Tunney, Ottawa (Ontario), K1A 0T6.
entreprises, 11 - Immeuble R.H. Coats, Statistique Canada, Parc Tunney, Ottawa (Ontario), K1A 0T6.

Niyonsenga présente une comparaison de deux méthodes non paramétriques d'estimation des probabilités de réponse dans la théorie de l'échantillonnage aléatoire simple sans remise, on peut constater que la variante non paramétrique basée sur les rangs des valeurs de la variable auxiliaire donne de meilleurs résultats, sur les plans du biais et de l'erreur quadratique moyenne, que la méthode basée sur les valeurs de la variable auxiliaire et ce, à la fois pour les estimateurs d'expansion et de régression.

Schabenberger et Gregoire comparent des stratégies π pt exacte et approximative dans le cas de l'échantillonnage pour les services forestiers. Ils comparent deux plans de sondage séquentiels de Sunter, combinés à l'estimateur d'Horvitz-Thompson, à la stratégie de groupe aléatoire de Rao, Hartley et Cochran (RHC), ainsi qu'à un estimateur par quotient de moyennes utilisé avec un sondage aléatoire simple. Si la taille de la variable est fortement corrélée avec la variable étudiée, alors les stratégies π pt sont considérablement plus efficaces. Quand la corrélation est très forte, la stratégie π pt est la plus efficace. Toutefois, la stratégie RHC présente l'avantage de la simplicité. Lorsque la corrélation est faible, les stratégies π pt peuvent être très inefficaces.

Le rédacteur en chef

Dans ce numéro

Au début de ce numéro de *Techniques d'enquête* se trouve une section spéciale consacrée aux méthodes d'enquête-établissements. Les quatre communications qu'elle contient traitent d'importantes questions, telles que la conception des questionnaires, les plans d'échantillonnage et l'estimation. Ces communications ont été présentées à l'origine à la conférence internationale sur les enquêtes-établissements qui a eu lieu à Buffalo (New York), en juin 1993.

Armstrong et St-Jean présentent une application du cadre général de l'estimation par régression dans les sondages à deux phases. En utilisant les données d'un échantillon à deux phases d'enregistrements fiscaux, ils comparent empiriquement trois cas particuliers de l'estimateur de régression généralisé, deux estimateurs de régression et un estimateur stratifié à posteriori, à l'estimateur d'Horvitz-Thompson. L'étude empirique révèle que l'estimateur stratifié à posteriori est plus efficace que celui d'Horvitz-Thompson, et qu'il est aussi efficace que les deux estimateurs de régression.

Gallego, Delinco et Carfagna décrivent le projet MARS (Monitoring Agriculture with Remote Sensing) de la Communauté européenne. Comme ce projet n'est pas en mesure de produire de bonnes estimations des surfaces cultivables et des rendements des récoltes, les auteurs décrivent une méthode d'échantillonnage par points des fermes afin d'obtenir des estimations fiables. Ils décrivent aussi les résultats de l'application de cette méthode à deux régions, Emilia Romagna en Italie et la République Tchèque. Pollock, Turner et Brown examinent l'utilisation de l'échantillonnage par saisie-ressaisie pour estimer la taille de la population et les totaux de la population lorsqu'il n'existe que des listes incomplètes. Ils examinent les propriétés des estimateurs par modèle obtenus et présentent un exemple qui utilise les navires de pêche.

Dans le dernier document de cette section spéciale, Gower présente une vue d'ensemble de considérations importantes lors du développement et de la conception de questionnaires pour les enquêtes-entreprises. Des exemples d'utilisation de groupes de discussion et de la recherche cognitive pour tester les questionnaires de ces enquêtes sont présentés. Rancourt, Lee et Särndal présentent des facteurs de correction simples pour réduire le biais de l'estimateur usuel de la moyenne de la population dans le cas d'une imputation par ratio pour la non-réponse confondu. On utilise des simulations de Monte Carlo pour étudier l'effet de ces facteurs. On a constaté que ces facteurs étaient efficaces, en particulier lorsque le modèle sous-jacent à l'imputation par quotient tient.

L'utilisation de la méthode de capture-recapture pour l'évaluation de la couverture du recensement des États-Unis fait l'objet de la communication de Ding et Fienberg. Les auteurs donnent des méthodes d'estimation du total de la population et du sous-dénombrement du recensement lorsqu'on relâche l'hypothèse d'un appariement parfait entre les individus du recensement et ceux de l'échantillon. Les auteurs proposent des modèles pour décrire deux types d'erreurs d'appariement, l'appariement erroné et le non-appariement erroné. Les méthodes sont illustrées à l'aide de données basées sur le recensement d'essai de 1986 de Los Angeles et le recensement décennal de 1990.

Kott discute le test d'une hypothèse à propos des coefficients de régression linéaire en utilisant des données provenant d'une enquête par sondage. Il propose une correction de l'estimateur linéarisé de la variance basé sur le plan d'échantillonnage pour réduire le biais dû au modèle et une formule afin d'estimer son nombre effectif de degrés de liberté. Il présente deux exemples. Cox développe un cadre, appelé masquage matriciel, pour les méthodes de limitation de la divulgation des micro-données, et qui devrait permettre de mieux comprendre ces méthodes et leurs effets sur l'utilisation des données. Utilisant ce cadre et le calcul matriciel ordinaire, les organismes statistiques peuvent développer, évaluer et utiliser des logiciels fiables pour limiter la divulgation des micro-données. L'auteur présente des formulations matricielles explicites pour les principales méthodes de masquage des micro-données utilisées actuellement.

Falorsi, Falorsi et Russo effectuent une comparaison empirique de quelques méthodes d'estimation de données régionales dans le cadre de l'Enquête sur la population active italienne, en prenant pour cela des données du recensement de l'Italie de 1981. Les estimateurs inclus dans leur étude sont un estimateur direct stratifié à posteriori, un estimateur synthétique, une combinaison linéaire optimale des deux et un estimateur dépendant de la taille de l'échantillon. Ils concluent que, pour leur application, ce dernier offre le meilleur équilibre de la variance et du biais.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada
Volume 20, numéro 2, décembre 1994

TABLE DES MATIÈRES

Dans ce numéro	99
Méthodes pour les enquêtes-établissements	
J. ARMSTRONG et H. ST-JEAN Estimation par régression généralisée pour un échantillon à deux phases de dossiers fiscaux	101
F.J. GALLEGO, J. DELINCE et E. CARFAGNA Échantillonnage à deux degrés dans des bases aréolaires sur des segments carrés pour des sondages agricoles	111
K.H. POLLOCK, S.C. TURNER et C.A. BROWN Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète	121
A.R. GOWER Conception des questionnaires d'enquêtes-entreprises	129
<hr/>	
E. RANCOURT, H. LEE et C.-E. SÄRNDALE Corrections du biais pour des estimations d'enquête tirées de données comprenant des valeurs imputées par quotient par suite d'une non-réponse selon un mécanisme confondu	143
Y. DING et S.E. FIENBERG Estimation de système dual du sous-dénombrement dans le recensement lorsqu'il y a erreur d'appariement	155
P.S. KOTT Test d'hypothèse portant sur des coefficients de régression linéaire et basé sur des données d'enquête	167
L.H. COX Méthodes de masquage de matrice pour la protection du caractère confidentiel de microdonnées	173
P.D. FALORSI, S. FALORSI et A. RUSSO Comparaison empirique de méthodes d'estimation pour petites régions pour l'enquête sur la population active italienne	179
T. NIYONSENGA Estimation non-paramétrique des probabilités de réponse en théorie de l'échantillonnage	185
O. SCHABENBERGER et T.G. GREGOIRE Solutions de remplacement pour les plans π pt authentiques: une étude comparative	193
Remerciements	201

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

B.N. Chinnappa

G.J.C. Hole

F. Mayda (Directeur de la Production)

R. Platek (Ancien président)

M.P. Singh

D. Roy

C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*

D. Binder, *Statistique Canada*

M.J. Colledge, *Australian Bureau of Statistics*

J.-C. Deville, *INSEE*

J.D. Drew, *Statistique Canada*

J.-J. Droesbeke, *Université Libre de Bruxelles*

W.A. Fuller, *Iowa State University*

M. Gonzalez, *U.S. Office of Management and Budget*

R.M. Groves, *University of Maryland*

D. Holt, *University of Southampton*

G. Kalton, *Westat, Inc.*

A. Mason, *East-West Center*

Rédacteurs adjoints

N. Laniel, M. Latouche, L. Mach et H. Mantel, *Statistique Canada*

A. Zaslavsky, *Harvard University*

K.M. Wolter, *National Opinion Research Center*

J. Waksberg, *Westat, Inc.*

J. Waite, *U.S. Bureau of the Census*

J. Sedransk, *State University of New York*

F.J. Scheuren, *George Washington University*

W.L. Schaible, *U.S. Bureau of Labor Statistics*

C.-E. Särndal, *Université de Montréal*

I. Sande, *Bell Communications Research, U.S.A.*

L.-P. Rivest, *Université Laval*

J.N.K. Rao, *Carleton University*

D. Pfeffermann, *Hebrew University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes-ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001 au catalogue) est de 45 \$ par année au Canada, 50 \$ (É.-U.) aux États-Unis, et de 55 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.



Ottawa

ISSN 0714-0045

N° 12-001 au catalogue

Autres pays : 55 \$ US

États-Unis : 50 \$ US

Prix : Canada : 45 \$

Décembre 1994

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

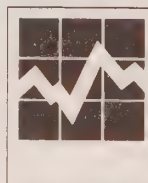
© Ministre de l'Industrie, des Sciences
et de la Technologie, 1994

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 1994 • VOLUME 20 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



20140 12



NUMÉRO 2

•

VOLUME 20

•

DÉCEMBRE 1994

PAR STATISTIQUE CANADA

ÉDITÉE

UNE REVUE

Colloque 18-PM1

TECHNIQUES D'ENQUÊTE

